Contributions to bias adjusted stepwise latent class modeling

Bakk, Zsuzsa

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

CONTRIBUTIONS TO BIAS ADJUSTED STEPWISE LATENT CLASS MODELING

Zsuzsa Bakk

CONTRIBUTIONS TO BIAS ADJUSTED STEPWISE LATENT CLASS MODELING

Zsuzsa Bakk
Tilburg University

CONTRIBUTIONS TO BIAS ADJUSTED STEPWISE LATENT CLASS MODELING

This research is funded by The Netherlands Organization for Scientific Research (NWO [VICI grant number 453-10-002]).

Printing was financially supported by Tilburg University.

ISBN: XXX
Printed by: Ridderprint BV, Ridderkerk, the Netherlands

CONTRIBUTIONS TO BIAS ADJUSTED STEPWISE LATENT CLASS MODELING

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
Tilburg University op gezag van de rector magnificus,
prof.dr. E.H.L. Aarts,
in het openbaar te verdedigen ten overstaan van
een door het college voor promoties aangewezen
commissie in
de aula van de Universiteit

op vrijdag 16 oktober 2015 om 14.15 uur

door

Zsuzsa Bakk

geboren op 16 mei 1982 te Targu Secuiesc, Roemenië

Promotor:	prof.dr. J. K. Vermunt
Copromotor:	dr. D.L. Oberski
Overige leden van de Promotiecommissie:	prof. F. Bassi
	dr. M.A. Croon
	dr. J.P.T.M. Gelissen
	prof.dr. P.G.M. van der Heijden
	prof.dr. J. Kuha

Contents

Motto

Klaarte is nie hier nie: klaarighede moontlik,
maar nie klaarte nie....Dis alles aan die word,
gedurigdeur.
(Clarity is not here: classification possibly,
not clarity....Everything still becomes
constantly)
Petra Müller: Gety (Tide)

Chapter 1

Introduction

1.1 Latent class modeling

Latent class analysis is an approach used in the social and behavioral sciences for classifying objects into a smaller number of unobserved groups (categories) based on their response pattern on a set of observed indicator variables. Examples of applications include the identification of types of political involvement (Hagenaars & Halman 1989), subgroups of juvenille offenders (Mulder, Vermunt, Brand, Bullens, & Van Marle, 2012), types of psychological contract (De Cuyper et al. 2008), and types of gender role attitudes (Yamaguchi 2000).

Identifying the unknown subgroups or clusters is usually just the first step in an analysis since researchers are often also interested in the causes and/or consequences of the cluster membership. In other words, they may wish to relate the latent variable to covariates and/or distal outcomes. For example, De Cuyper et al. (2008) investigated whether being on a temporary or permanent contract has an impact on the type of psychological contract that exists between the employee and employer (relating LC membership to covariates), as well as whether the type of psychological contract has an impact on job and life satisfaction, organizational commitment, and contract violation (relating LC membership to distal outcomes). Similarly, not only identifying groups of juvenile offenders is important, but also seeing their recidivism pattern, a research question that in the work of Mulder et al. meant exploring the relationship between LC membership with more than 70 distal outcomes.

Until recently there were two possible ways to relate LC membership to external variables of interest, namely, the one-step or the three-step approach presented in the following. Let us denote the latent class variable by X, the vector of indicators by $\mathbf{Y}$, the covariate (predictor of LC membership) by Z_p, and the distal outcome by Z_o. While throughout this chapter for simplicity we refer to a single external variable, both the covariate and distal outcome could be a vector of variables.

Using the one-step approach, the relation between the external variables Z_p and/or Z_o and the latent class variable is estimated simultaneously with the measurement model defining the latent classes (Dayton & Macready 1988; Hagenaars 1990; Yamaguchi 2000;

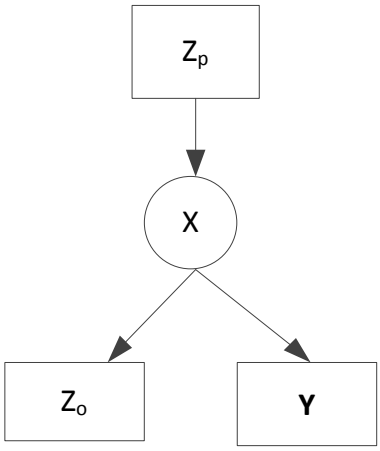

Figure 1.1: Associations between the latent variable (X), its indicators (Y), and external variables (Z) which can be outcome variables (Z_o) or predictor variables (Z_p).

Muthen 2004), as is shown in the model depicted in Figure 1.1. While Figure 1.1 shows the simplest association structure, a more complex model may also include direct effects of covariates on distal outcomes and/or indicator variables, as well as associations between distal outcomes and indicators.

The one-step approach is hardly ever used by practitioners, mostly because of the reasons enumerated below.

I Researchers prefer to separate the measurement part (relating the latent variable to the indicators) and the structural part (relating the latent variable to the external variables of interest) of the model especially when more complex models are investigated.

II When LC membership is related to a distal outcome using the one-step approach, this later is added to the LC model as an additional indicator. This means that unwanted assumptions need to be made about the conditional distribution of the distal outcome given the latent variable.

III Furthermore, an unintended circularity is created: while the interest is in explaining the distal outcome by the LC membership, the distal outcome contributes to the formation of the latent classes.

Until recently the only alternative to the one-step approach was the three-step approach. As depicted in Figure 1.2, when using this approach, first the underlying latent

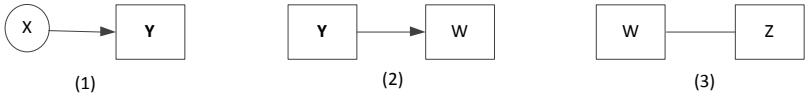

Figure 1.2: The steps of the standard three-step approach

class variable (X) is identified based on a set of observed indicator variables ($\mathbf{Y}$), then individuals are assigned to latent classes (we denote the class assignments by W), and subsequently the class assignments are used in further analyses investigating the W-Z relationships (Hagenaars, 1990). This approach tackles problem I, since the measurement and structural part of the model are separated. However, this approach also has an important deficit, namely, that the classification error introduced in the second step is ignored. This leads to biased estimates of the association of LC membership and external variables (Hagenaars, 1990; Bolck, Croon, and Hagenaars, 2004).

1.2 Bias adjusted stepwise LC models

Bolck, Croon and Hagenaars (2004) showed that the amount of classification error introduced in step two can be estimated and accounted for in the step-three analyses. These authors show that the true score on X can be re-obtained in step three by weighting W by the inverse of the classification errors. The approach, which we refer to as the BCH approach, proceeds as follows: the data on covariates and the classification are summarized in a multidimensional frequency table, the cell frequencies are reweighted by the inverse of the classification error matrix, and lastly a logit model is estimated using the reweighted frequency table as data, which yields the log-odds ratios describing the relationship between the external variables and the class membership. It should be mentioned that a similar approach was proposed by Fuller (1987), however has not been implemented.

The BCH approach is general, in the sense that it can be used in any situation that boils down to estimating the log-odds ratios in a contingency table, thus can be used with both covariates and distal outcomes as long as these are categorical variables. While the BCH approach offers a breakthrough by highlighting that the amount classification error is estimable and can be accounted for, it also has various disadvantages. That is, it can be used with categorical variables only, it is somewhat tedious since a new reweighted frequency table has to be created for each set of external variables, and it yields standard errors which are severely downward biased.

Vermunt (2010) suggested one important modification of the BCH approach, which eliminates the abovementioned limitations. Instead of creating and analyzing reweighted frequency tables, he proposed creating an expanded data file with T records per individual, where T is the number of latent classes. An additional column contains the BCH weights for each individual-class combination. The step-three model of interest can subsequently

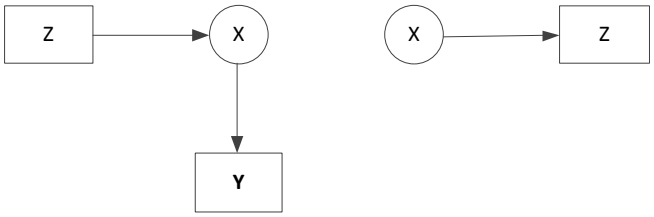

Figure 1.3: The steps of the LTB approach

be estimated using pseudo maximum likelihood methods, where the BCH weights are used as sampling weights. With this extended BCH approach, the latent class variable can be related to continuous covariates as well. Moreover, the bias in the standard errors (SEs) can be prevented by using a sandwich estimator that accounts for the weighting and the clustering in the expanded data file. When referring to the BCH approach in the remainder of this text, we mean this amended version, which is also the one which is currently used in practice.

Vermunt also proposed an alternative more direct bias-adjusted three-step approach, which he called the ML approach. It involves estimating a LC model in step three, with W as the single indicator variable having known classification error probabilities. Thus, while the BCH method weights W by the inverse of the classification error probabilities in a model for observed variables only, the ML approach estimates a LC model using the classification error probabilities as fixed (known) parts of the model, and freely estimates the structural part of the model in which LC membership is predicted by covariates.

A few unsolved problems with the ML and amended BCH approaches are that they can be used only with models with covariates, and the SE estimates are still somewhat downward biased. The reason for this bias is that in the step three model the estimates from step one are used as known values, while they are estimates having sampling fluctuation.

Another stepwise approach recently proposed specifically for models with distal outcomes is the LTB approach, so named after the developers, Lanza, Tan and Bray (2013). This approach was specifically developed to tackle the problem of the one-step approach presented above, namely that assumptions need to be made about the conditional distribution of the outcome(s) given the classes. This LTB approach is a two-step method in which first a LC model is estimated in which the distal outcome is used as a covariate in a one-step estimation procedure (see Figure 1.3). Using the outcome as covariate affecting LC membership no distributional assumptions are made about the outcome. In the second step, the class-specific means of the distal outcome are calculated using the model parameters obtained in the first step. A few problems of this approach are that the SE estimators available in literature are strongly downward biased, and using the approach with multiple distal outcomes is not well developed.

In summary, we can say that in the recent years various important improvements

have been proposed to bias-adjusted stepwise latent class modeling. Nevertheless, the ML, BCH, and LTB approaches are rather new, and still much is unknown about their performance under different circumstances. Furthermore, the approaches still have certain limitations, such as that the three-step approaches (BCH and ML) can be used only with covariates and that the LTB approach can deal only with a single distal outcome.

1.3 Outline of the thesis

This thesis proposes to contribute to the development of bias-adjusted stepwise modeling in three main aspects:

1. extend the ML and amended BCH approaches to models with distal outcomes and multiple latent variables;
2. amend for the bias in the SE estimates of the ML method that are caused by not accounting for the uncertainty about the fixed parameters;
3. analyze the robustness of the ML, BCH, and LTB approaches when applied with continuous distal outcomes, and present three possible improvements of the LTB approach.

In Chapter 2 we show how the ML and amended BCH approaches can be extended to a wider range of models. We show how the correction developed for the conditional distribution of the LC variable given the covariates can be generalized to modeling the joint distribution of class membership and external variables, from where specific subcases can be derived. For example in case of relating LC membership to a distal outcome using the BCH approach a weighted ANOVA is performed, while with the ML approach a LC model is estimated with 2 indicators: W and Z, where the misclassification probabilities for W are assumed to be known. We show that as long as all model assumptions hold both the ML and BCH approaches are unbiased estimators of the association between LC membership and distal outcomes or of the association between multiple LC variables.

Next in Chapter 3 we pay attention to the SE estimators of the ML approach. While the parameter estimates obtained with this approach are unbiased, there is still some bias left in the SE estimates that is due to ignoring the sampling fluctuation of the fixed value parameters. We propose investigating several candidate SE estimators that can account for this additional source of uncertainty based on the literature on non-linear models (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006), three-step structural equation modeling (Skrondal & Kuha, 2012; Oberski & Satorra, 2013), and econometric theory for two-stages least squares (Murphy & Topel, 1985). We apply the general theory of Gong and Samaniego (1981) to latent class modeling, noting similarities and differences with these other approaches.

Furthermore in Chapter 4 we investigate the robustness of the ML, BCH and LTB approaches when applied to models with continuous distal outcomes. Note that while the LTB approach was specifically developed for these type of models, the use of the ML and BCH approach for these models was proposed only in Chapter 2 of this dissertation. While

all three approaches perform well when the underlying model assumptions hold, we can expect that some of the approaches are less robust for violations of these assumptions. We can expect that the BCH approach, that is an ANOVA is more robust than the ML approach to violations of normality. At the same time the LTB approach assumes that the relationship between the continuous outcome variable and the LC membership is linear-logistic. The impact of the violation of this assumption on the class-specific means calculated in step two is unknown.

Based on the results of Chapter 4 we recommend a few extensions to the LTB approach in Chapter 5. First in the spirit of this dissertation, a true stepwise implementation is provided in which the building of the latent classes and the investigation of the relationship of the classes with the distal outcomes is separated. This simplifies the analysis in situations where the LC membership should be related to multiple distal outcomes. As a second extension, similar to quadratic discriminant analysis, the inclusion of a quadratic term in the logistic model for the LCs is proposed, for situations where the variances of the continuous distal outcome differs across LCs, thus violating the assumption of linear-logistic association. The quadratic term prevents that one obtains biased estimates of the class-specific means in such situations. The third extension involves estimating the standard errors of the class-specific means by means of jackknife or a (non-parametric) bootstrap procedure. Both SE estimators proposed here yield much better coverage rates than the currently available estimator which shows clear undercoverage.

Chapter 2

Estimating the association between latent class membership and external variables using bias adjusted three-step approaches

Abstract

Latent class (LC) analysis is a clustering method widely used in social science research.Usually the interest lies in relating the clustering to external variables. This can be done using a three-step approach, which proceeds as follows: the LC model is estimated (step 1), predictions for the class membership scores are obtained (step 2) and used to assess the relationship between class membership and other variables (step 3). Bolck, Croon, and Hagenaars (2004) showed that this approach leads to severely biased estimates of the third step estimates, and proposed correction methods, that were further developed by Vermunt (2010). In the current study, we extend these correction methods to situations where class membership is not predicted but used as an explanatory variable in the third step. A simulation study tests the performance of the proposed correction methods, and their practical use was illustrated with real data examples. The results show that the proposed correction methods perform well under conditions encountered in practice.

This chapter is published as Bakk, Z., Tekle, F.B. & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias adjusted three-step approaches. *Sociological Methodology, vol.43, 1 pp. 272-311*

2.1 Introduction

The use of latent class analysis (LCA) (Lazarsfeld & Henry, 1968; Goodman, 1974; McCutcheon, 1987) is becoming more and more widespread in social science research, especially because of increasing modeling options and software availability. In its basic form, LCA is a statistical method for grouping units of analysis into clusters, that is, to identify subgroups that have similar values on a set of observed indicator variables. Examples of applications include the identification of types of political involvement (Hagenaars & Halman 1989), types of psychological contract (De Cuyper et al. 2008), types of gender role attitudes (Yamaguchi, 2000), and types of music consumers (Chan & Goldthorpe 2007).

Identifying the unknown subgroups or clusters is usually just the first step in an analysis since researchers are often also interested in the causes and/or consequences of the cluster membership. In other words, they may wish to relate the latent variable to covariates and distal outcomes. There are two possible ways to proceed with this latter extension, namely, using a one-step or a three-step approach. Using the one-step approach, the relation between the external variables of interest (covariates and/or distal outcomes) and the latent class variable is estimated simultaneously with the model for identifying the latent variable (Dayton & Macready 1988; Hagenaars 1990; Yamaguchi 2000; Van der Heijden, Dessens & Bockenholt 1996). Using the other alternative, the three-step approach, first the underlying latent construct is identified based on a set of observed indicator variables, then individuals are assigned to latent classes, and subsequently the class assignments are used in further analyses (Bolck et al. 2004; Vermunt 2010). When all the model assumptions hold, the more complex one-step approach is better from a statistical point of view, because it is more efficient.

However, most applied researchers prefer using the simpler three-step approach. De Cuyper et al. (2008) and Chan & Goldthorpe (2007) use such a three-step approach with covariates, as do Olino et al. (2011) with distal outcomes. One reason for using the three-step approach is that researchers see constructing a latent typology and investigating how the latent typology is related to external variables as two different steps in an analysis. For instance, in an LCA with distal outcomes, the latent classes will typically be risk groups (e.g., groups of youth delinquents based on delinquency histories or groups of persons with different lifestyles), and the distal outcomes are events in a later life stage (e.g., recidivism or health status). It is substantively difficult to argue that the distal outcomes should be included in the same model as the one that is used to identify the risk groups if one wishes to investigate the predictive validity of the latent classification.

Another argument for the three-step approach as opposed to the one-step is that in applications wherein a possibly large set of external variables is considered, the estimation procedure for the latter approach might fail because of the sparseness of the analyzed frequency table and the potentially large number of parameters (Goetghebeur, Liinev, & Boelaert, 2000; Huang & Bandeen-Roche, 2004; Clark & Muthen, 2009). For example, in a study by Mulder et al. (2012), the association of subgroups of recidivism with 70 possible distal outcomes was analyzed, which would be impossible using the one-step approach.

A related problem with the one-step approach is that the inclusion of covariates or

distal outcomes can distort the class solution because additional assumptions are made that may be violated (Huang, Brecht, Hara, & Hser, 2010; Tofighi & Enders, 2008; Bauer & Curran, 2003; Petras & Masyn, 2010). For example, the inclusion of a distal outcome requires specification of its within-class distribution, which if misspecified can distort the whole class solution. It may even happen that rather different class solutions are obtained when different distal outcomes are included separately in the model, though theoretically the latent classes should be based on the indicators and predict only the distal outcome.

Although there are many situations in which researchers may prefer the three-step LCA, the main disadvantage of this approach is that it yields severely downward-biased estimates of the association between class membership and external variables (Bolck et al. 2004; Vermunt 2010). Recently, several correction methods were developed to tackle this problem. Clark and Muthen (2009) proposed a correction method based on pseudo class draws from their posterior distribution. However this approach, still maintains a relatively large bias in the log odds ratios of the association of the latent class variable with covariates. Petersen, Bandeen-Roche, Budtz-Jrgensen, and Groes (2012) developed a method based on a translation of the idea of Bartlett scores to the LCA context, which in the simulation study performed by the authors turned out to perform well. Bolck et al. (2004) developed a correction method that involves analyzing a reweighted frequency table and that can be used in three-step LCA with categorical covariates. Later Vermunt (2010) suggested a modification of this method, making it possible to obtain correct standard errors (SEs) and accommodate continuous covariates, and also introduced a more direct maximum likelihood (ML) correction method.

A limitation of the currently available adjustment methods for three-step LCA is that they were all developed and tested for the situation wherein class membership is treated as depending on the external variables. Moreover, all these methods were studied using models with only a single latent variable. However, applied researchers are often interested in a much broader use of the latent class solution. Therefor there should be correction methods available for a larger variety of modeling options. Given this gap in the literature, in the current article, we show how the three-step correction methods developed by Bolck et al. (2004) and Vermunt (2010) can be adapted to the situation in which the latent variable is a predictor of one or more distal outcomes, which may be categorical or continuous variables. We also pay attention to the situation in which the distal outcome itself is also a categorical latent variable. This implies that one should adjust for classification errors in both the predictor and the outcome variable.

The content of the article is outlined as follows. First we introduce the basic latent class model and discuss class assignment and quantification of the associated classification error. Then, the two classic ways of handling external variables in LCA will be presented (namely, the one-step and three-step approaches). Next, we discuss the correction methods developed by Bolck et al. (2004) and Vermunt (2010) for three-step LCA and show how these can be generalized for modeling the joint distribution of class membership and external variables, from where specific subcases can be derived. Subsequently, we check the performance of the different correction methods using a simulation study and illustrate them with real data applications.

2.2 Latent class modeling and classification

2.2.1 The basic latent class model

Let us denote the categorical latent variable by X, a particular latent class by t, and the number of classes by T, as such we have $t = 1, 2, ...T$. Let Y_k represent one of the K manifest indicator variables, where $k = 1, 2, ...K$. Let $\mathbf{Y}$ be a vector containing a full response pattern and $\mathbf{y}$ its realization. A latent class model for the probability of observing response pattern $\mathbf{y}$ can be defined as follows:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{t=1}^{T} P(X = t)P(\mathbf{Y} = \mathbf{y}|X = t), \tag{2.1}$$

where $P(X = t)$ represents the probability of belonging to class t and $P(\mathbf{Y} = \mathbf{y}|X = t)$ the probability of having response pattern $\mathbf{y}$ conditional on belonging to class t. As we can see from Equation 2.1, the marginal probability of obtaining response pattern $\mathbf{y}$ is assumed to be a weighted average of the t class-specific probabilities.

In a classical LCA we assume local independence, which means that the K indicator variables are assumed to be mutually independent within each class t. This implies that, the joint probability of a specific response pattern on the vector of indicator variables is the product of the item specific probabilities:

$$P(\mathbf{Y} = \mathbf{y}|X = t) = \prod_{k=1}^{K} P(Y_k|X = t), \tag{2.2}$$

Combining Equation 2.1 and 2.2 we obtain the following:

$$P(\mathbf{Y}) = \sum_{t=1}^{T} P(X = t) \prod_{k=1}^{K} P(Y_k|X = t). \tag{2.3}$$

The model parameters of interest are the class proportions $P(X = t)$ and the class-specific response probabilities $P(\mathbf{Y} = \mathbf{y}|X = t)$. These parameters are usually estimated by maximum likelihood (ML).

2.2.2 Obtaining latent class predictions

While the true class memberships cannot be observed, the parameters of the measurement model described in Equations 2.1 to 2.3 can be used to derive procedures for estimating these class memberships, that is, for assigning individuals to classes (Goodman 1974, 2007; Hagenaars 1990). The prediction is based on the posterior probability of belonging to class t given an observed response pattern $\mathbf{y}$, $P(X = t|\mathbf{Y} = \mathbf{y})$, which can be obtained by using Bayes' theorem, that is:

$$P(X = t|\mathbf{Y} = \mathbf{y}) = \frac{P(X = t)P(\mathbf{Y} = \mathbf{y})}{P(\mathbf{Y} = \mathbf{y})}. \tag{2.4}$$

These posterior class membership probabilities provide information about the distribution over the T classes among individuals with response pattern $\mathbf{y}$, which reflects that persons having the same response pattern can belong to different classes. It is important to note that each individual belongs to only one class but that we do not know to which. Using the posterior class membership probabilities, different types of rules can be used for assigning subjects to classes, the most popular of which are modal and proportional assignment. When using modal assignment, each individual is assigned to the class for which its posterior membership probability is the largest. Denoting the predicted class by W and subject $i^{'}$s response pattern by $\mathbf{y}_i$, the hard partitioning corresponding to modal assignment can be expressed as the following:

$$P(W = s|\mathbf{Y} = \mathbf{y}_i) = \begin{cases} 1 \text{ if } P(X = s|\mathbf{Y} = \mathbf{y}_i) > P(X = t|\mathbf{Y} = \mathbf{y}_i) \forall s \neq t. \\ 0 \text{ else.} \end{cases}$$

An individual is assigned with probability or weight equal to 1 to the class with the largest posterior probability and with weight 0 to the other classes. Below we will also use the shorthand notation w_{is} for $P(W = s|\mathbf{Y} = \mathbf{y_i})$.

To illustrate the class assignment, let us assume that we have a two-class model and that for a particular response pattern containing 20 respondents we find a probability of 0.8 of belonging to class 1, and of 0.2 of belonging to class 2. This means that 16 persons belong to class 1 and 4 to class 2. Under modal assignment, all 20 individuals will be assigned to class 1, which means that 4 will be misclassified (but we do not know who). This can be expressed as follows: 16*(0) + 4*(1) = 4. It should be noted that modal assignment is optimal in the sense that the number of classification errors is smaller than with any other assignment rule.

An alternative to modal assignment is proportional assignment, which in the context of model-based clustering is referred to as a soft partitioning method (Dias and Vermunt 2008). An individual with the response pattern $\mathbf{y}_i$ will then be assigned to each class s with a weight $P(W = s|\mathbf{Y} = \mathbf{y}_i) = P(X = s|\mathbf{Y} = \mathbf{y}_i)$. That is, with a weight equal to the posterior membership probability. In our example, this would mean that each of the 20 observations receive weights of .8 and .2 for belonging to the first and second class, respectively. In practice, this is achieved by creating an expanded data file with one record per class per respondent and by using the class membership probabilities as weights in subsequent analyses.

While at first glance it may seem that proportional assignment prevents introducing misclassifications, this is clearly not the case. In our example, the 16 persons belonging to class 1 receive a weight of .8 for class 1 instead of a weight of 1, which corresponds to a misclassification of .2, and the 4 persons belonging to class 2 receive a weight of .2 for class 2 instead of a weight of 1, which corresponds to a misclassification of .8. The total number of misclassifications for the data pattern concerned is therefore 16*(.2) + 4*(.8) = 6.4.

Although modal and proportional assignment are the most common methods, it is also possible to use other rules. An example is the random assignment of individuals to classes based on the posterior class membership probabilities, which is in fact a stochastic version of the proportional assignment rule. The expected number of misclassification is the same

under random and proportional assignment. A rule similar to modal assignment involves assigning individuals to class s if the posterior probability is larger than a threshold. For example, in a two class model, one assigns an individual to class 1 if the posterior membership probability for this class is larger than .7 and otherwise to class 2. Compared to modal assignment, such a rule reduces the number of misclassifications into class 1 but increases the misclassifications into class 2.

It is clear that irrespective of the assignment method used, class assignments and true class scores will differ for some individuals (Hagenaars 1990; Bolck et al. 2004). As is shown in more detail below, the overall proportion of misclassifications can be obtained by averaging the misclassification probabilities of all data patterns. This overall classification error can be calculated irrespective of the assignment rule applied.

2.2.3 Quantifying the classification errors

The overall quality of the classification obtained from a LCA can be quantified by $P(W = s|X = t)$; that is, by the probability of a certain class assignment conditional on the true class. The larger the probabilities for $s = t$, the better the classification. Using the LCA parameters this quantity can be obtained as follows [2]:

$$\begin{aligned} P(W = s|X = t) &= \sum_{Y} P(\mathbf{Y} = \mathbf{y}|X = t)P(W = s|\mathbf{Y} = \mathbf{y}) \\ &= \sum_{Y} \frac{P(\mathbf{Y} = \mathbf{y})P(X = t|\mathbf{Y} = \mathbf{y})P(W = s|\mathbf{Y} = \mathbf{y}}{P(X = t)}. \end{aligned} \tag{2.5}$$

In fact, the overall classification errors are obtained by averaging the classification errors for all possible response patterns. As indicated by Vermunt (2010), when the possible number of response patterns is very large, it is more convenient to estimate the classification errors by averaging over the patterns occurring in the sample, which involves replacing $P(\mathbf{Y} = \mathbf{y})$ by its empirical distribution:

$$P(W = s|X = t) = \frac{\frac{1}{N}\sum_{i=1}^{N} P(X = t|\mathbf{Y} = \mathbf{y}_i)w_{is}}{P(X = t)}, \tag{2.6}$$

where N is the sample size and as indicated above $w_{is} = P(W = s|\mathbf{Y} = \mathbf{y}_i)$. Below we will show how $P(W = s|X = t)$ is used in the correction methods for three-step LCA.

The concept of classification error is strongly related to the concept of separation between classes. The latter refers to how well the classes can be distinguished based on the available information on $\mathbf{Y}$. More specifically, lower separation between classes corresponds to larger classification errors. Measures for class separation, and therefor also for classification error, quantify how much the posterior membership probabilities $P(X = s|\mathbf{Y} = \mathbf{y}_i)$ deviate from uniform. For this purpose, one can (among others) use

[2] Note that in Equation 2.5, we implicitly use the equality $P(W|Y, X) = P(W|Y)$. This follows from the fact that class assignment depends only on Y (and the latent class analysis model parameters) but not directly on X.

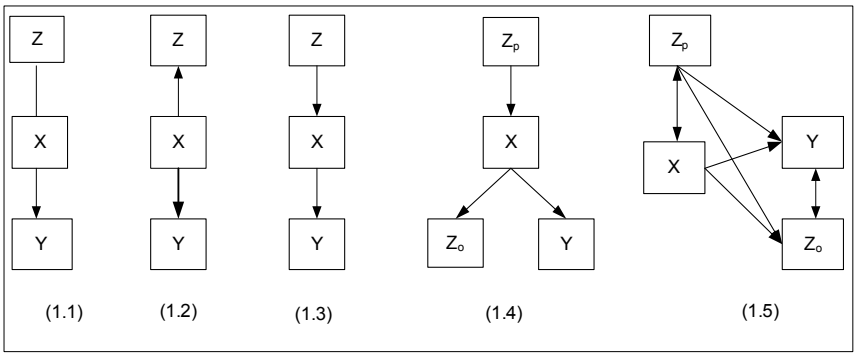

Figure 2.1: Types of associations between the latent variable (X), its indicators (Y), and other external variables (Z) that can be outcome variables (Z_o) or predictor variables (Z_p) of the latent variable.

the principle of entropy: $-\sum_{t=1}^{T} P(X=t|\mathbf{Y}=\mathbf{y})\log P(X=t|\mathbf{Y}=\mathbf{y})$. The proportional reduction of entropy when $\mathbf{Y}$ is available compared to the situation in which $\mathbf{Y}$ is unknown is a pseudo R^2 measure for class separation (Vermunt & Magidson, 2013), and thus also for the quality of the classification of a sample.

2.3 LCA with external variables: traditional approaches

There are a variety of ways in which external variables may play a role in a LCA; the most common ones are depicted in Figure 2.1(2.1.1 - 2.1.5). We denote an external variable by Z, the latent variable by X, and the vector of indicators by $\mathbf{Y}$. It should be noted that while the use of multiple latent variables is possible, for clarity of exposition, in the main part of the current paper, we focus on the situation of a single X and illustrate the possibility of extension to multiple latent variables in one of the empirical examples.

In its most general form, we can think of the latent class variable X being measured by its indicators $\mathbf{Y}$ and being associated with external variables Z, without specifying a causal order between X and Z (Figure 2.1.1). More specific cases are when Z is a distal outcome (Figure 2.1.2), when Z is a predictor of X (Figure 2.1.3), or when Z contains both predictors Z_p and distal outcomes Z_o (Figure 2.1.4). The most general form of an association between X and Z, without specifying a causal order (Figure 2.1.1) involves modeling the joint probability of the three sets of variables as follows:

$$P(Z=z, X=t, \mathbf{Y}=\mathbf{y}) = P(Z=z, X=t)P(\mathbf{Y}=\mathbf{y}|X=t). \tag{2.7}$$

Note that in this expression we make the assumption that Z and $\mathbf{Y}$ are conditionally independent of one another given X. This means that Z is associated with X, but controlling for X it is not associated with the indicators. This is a rather standard

assumption in latent variables models with external variables, which is moreover needed for the adjusted three-step approaches.

Based on the substantive theoretical arguments about the causal relationship between X and Z, the joint distribution in Equation 2.7 can be adapted to accommodate specific cases. For instance, if we assume that the latent variable depends on the external variable, the relationship between X and Z can be analyzed using a model of the form (see Figure 2.1.3):

$$P(Z = z, X = t, \mathbf{Y} = \mathbf{y}) = P(Z = z)P(X = t|Z = z)P(\mathbf{Y} = \mathbf{y}|X = t).$$

Because the marginal distribution of Z is typically not of interest, it can be dropped and the model can be defined as follows:

$$P(X = t, \mathbf{Y} = \mathbf{y}|Z = z) = P(X = t|Z = z)P(\mathbf{Y} = \mathbf{y}|X = t). \tag{2.8}$$

Another type of situation that is often of interest is when the latent variable is a predictor of the external variable (see Figure 2.1.2). In this case, we use a model of the form [3]:

$$P(Z = z, X = t, \mathbf{Y} = \mathbf{y}) = P(X = t)P(Z = z|X = t)P(\mathbf{Y} = \mathbf{y}|X = t). \tag{2.9}$$

When some of the Z variables are predictors and others outcomes (Figure 2.1.4), the model becomes:

$$\begin{aligned} P(Z_p = z_p, X = t, \mathbf{Y} = \mathbf{y}, Z_o = z_o) = P(X = t|Z_p = z_p) \\ P(Z_o = z_o|X = t, Z_p = z_p)P(\mathbf{Y} = \mathbf{y}|X = t) \end{aligned}$$

where Z_o is the distal outcome variable, and Z_p a covariate. Note that the latter two models require the specification of the conditional distribution of Z (Z_o) in order to quantify the effect of X on Z. In the current paper, we will use a normal distribution for continuous Z and a multinomial distribution for ordinal and nominal Z. The regression models used are linear, cumulative logistic, and multinomial logistic regression (Agresti 2002).

When the implied conditional independence assumption holds, each of the four variants described above can be investigated using either a one-step or a three-step procedure. However when this is not the case, one may prefer using a one-step approach, in which it is possible to relax the assumption that Z and $\mathbf{Y}$ are conditionally independent given X (Huang and Bandeen-Roche 2004), contrary to the three step approaches where this is not yet possible [4]. Extensions of the standard latent class model using the one-step approach

[3] While in Equation 2.8 it is clear that the extension to more covariates Z is straightforward, this is also possible using Equation 2.9, assuming conditional independence of outcomes given X.

[4] Although in this article we emphasize the need of the conditional independence assumption to hold to be able to use any of the three-step methods, it should be mentioned that an extension of the corrected three-step approaches could be developed that makes it possible to include direct effects of categorical covariates on indicators in the model. This could be done by applying the weighting that we present in the following pages separately at every level of the external covariate.

make it possible to include direct effects of covariates on indicators, or residual correlations between indicators and distal outcomes, as shown on Figure 2.1.5. Readers interested in such extensions are referred to the literature available on these models (Hagenaars, 1988; Bandeen-Roche, Miglioretti, Zegger, & Rathouz, 1997; Huang & Bandeen-Roche, 2004) . It should be mentioned that when the assumptions of conditional independence of Z and $\mathbf{Y}$ is violated this can influence model parameters. There is a need to further investigate whether the three or the one-step approach is more affected by this problem.

In the following we will restrict ourselves to the situation in which Z and $\mathbf{Y}$ can be assumed to be independent given X. We will show how the relevant models can be estimated using one-step LCA, standard three-step LCA, and bias adjusted three-step LCA.

2.3.1 One-step approach

Using this approach, the external variables are incorporated in the latent class model and the resulting extended model is estimated simultaneously with the measurement model. The extended model can be seen as being composed of two parts: the measurement model that comprises information on $\mathbf{Y}$ given X, and the structural part that deals with the relationship between X and Z.

Both covariates (Figure 2.1.3) and distal outcome variables (Figure 2.1.2) can be included, possibly in combination with one another (Figure 2.1.4). In situations where the class membership is used as a predictor of one or more external distal outcomes Z, the latter have a role similar to those of the indicator variables (Hagenaars 1990:135-142; Huang et.al 2010).

2.3.2 The standard three-step approach

The method which is presented graphically in Figure 2.2 proceeds as follows. In the first step, the measurement model for the relationship between the latent variable and its indicators is built, as described by Equation 2.3 and depicted in Figure 2.2.1. In the next step, using the information from the first step, subjects are assigned to latent classes based on their scores on the indicator variables, as depicted in Figure 2.2.2. In this process different assignment rules can be used, the most common ones being modal and proportional assignment. In the third step, the predicted class membership variable (W) is used in further analysis, implying analyzing the relationship between W and Z (Figure 2.2.3).

Bolck et al. (2004) proved that the estimates of the log-odds ratios characterizing the relationship between Z and W will always be smaller than those characterizing the relationship between Z and X, and proposed a correction method that can be used with categorical external predictors (Figure 2.1.3). Their correction method was later extended by Vermunt (2010), who showed how to adjust for the downward bias in the standard errors (SE) obtained by the initial method and how to include continuous covariates in the step-three model. Vermunt (2010) also proposed a maximum likelihood (ML) based correction method. In the following, we present these two correction methods and show

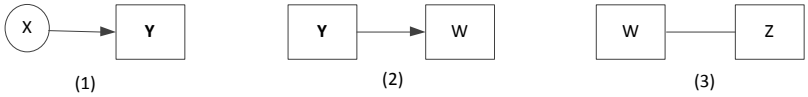

Figure 2.2: The steps of the standard three-step approach

how these can be generalized to the situation in which the class membership is a predictor instead of an outcome variable.

2.4 Generalization of existing correction methods

While in the standard three-step procedure we estimate the relationship between W and Z, actually we are interested in the relationship between X and Z. The key to the correction methods lies in the fact that it is possible to show how the $X-Z$ distribution is related to the $W-Z$ distribution. Let us first refer to Figure 2.3, which shows how the four (sets) of variables of interest are connected. From the joint distribution of X, Z, W, and $\mathbf{Y}$, we can derive the marginal distribution of W and Z by summing over all possible values of X and $\mathbf{Y}$; that is,

$$\begin{aligned} P(W=s,Z=t) &= \sum_t \sum_y P(X=t,Z=z,\mathbf{Y}=\mathbf{y},W=s) \\ &= \sum_t P(X=t,Z=z) \sum_y P(\mathbf{Y}=\mathbf{y},W=s|X=t,Z=z) \\ = \sum_t P(X=t,Z=z) \sum_y P(\mathbf{Y}=\mathbf{y}|X=t,Z=z) & P(W=s|X=t,Z=z,\mathbf{Y}=\mathbf{y}). \end{aligned}$$

Given that W depends only on $\mathbf{Y}$ (as a consequence of the way the class assignment are obtained), and assuming that Z is independent of $\mathbf{Y}$ given X (the assumption depicted in Figure 2.1.1), and subsequently replacing $P(\mathbf{Y}=\mathbf{y}|X=t)$ by $(P(\mathbf{Y}=\mathbf{y})P(X=t|\mathbf{Y}=\mathbf{y}))/P(X=t)$ using Bayes theorem we obtain:

$$\begin{aligned} P(W=s,Z=z) &= \sum_t P(X=t,Z=z) \\ &\frac{\sum_y P(\mathbf{Y}=\mathbf{y})P(X=t|\mathbf{Y}=\mathbf{y})P(W=s|\mathbf{Y}=\mathbf{y})}{P(X=t)} \\ &= \sum_t P(X=t,Z=z)P(W=s|X=t). \end{aligned} \tag{2.10}$$

The last substitution follows from the definition presented in Equation 2.5. As can be seen from Equation 2.10, the entries in the W and Z distribution are weighted sums

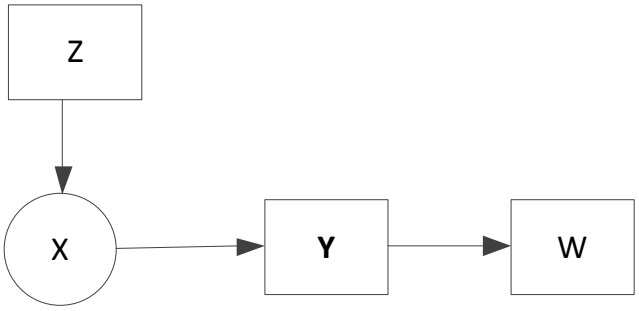

Figure 2.3: The relationship between variables W, X, Y and Z in the three-step approach.

of the entries in the X and Z distribution, where the weights are the misclassification probabilities $P(W = s|X = t)$. This suggests that the relationship between X and Z can be obtained by adjusting the relationship between W and Z for the misclassification probabilities $P(W = s|X = t)$.

The correction methods developed by Bolck, et al. (2004) and Vermunt (2010) are based on an equality similar to the one described in Equation 2.10. The difference is that these concern the relationship between the conditional distributions of X given Z and W given Z, so the situation where Z is a covariate and X is the outcome. As we have shown above in Equation 2.10, the correction methods can also be applied to the joint distribution of X and Z. From this joint distribution the conditional distribution of Z given X can be obtained when the latent variable X is considered to be a predictor of external variable Z. The extension of the methods lies on the realization that the classification error depends only on the measurement model. The consequence of this is that irrespective of the role of X and Z in describing their mutual relationship, the adjustments remain the same. The same type of adjustments can also be used with multiple latent variables as we will discuss shortly in a later section.

2.4.1 The three-step ML approach

The ML-based correction method introduced by Vermunt (2010) involves defining a latent class model with one or more covariates Z affecting the latent variable X and with the predicted class membership W as the single indicator of the underlying latent variable X. An important difference compared to a standard LCA is that the conditional response probabilities $P(W = s|X = t)$ are not estimated but fixed to their estimated values from the previous step.

Vermunt's procedure can easily be adapted for the modeling of the joint distribution of X and Z or the conditional distribution of Z given X. As can be seen from Equation 2.10, even if we have information only on Z and W and if $P(W = s|X = t)$ is known, it is possible to specify a (latent class) model yielding information on the association between X and Z. This requires using W as an indicator of X and defining the form of the $X - Z$ distributions. Equation 2.10 can also be re-expressed as follows:

$$P(W = s, Z = z) = \sum_t P(X = t)P(Z = z|X = t)P(W = s|X = t), \tag{2.11}$$

corresponding to the situation in which X is a predictor of Z. Note that this yields a latent class model with two indicators, Z and W, where W comprises all the information on the classification from the first two steps. An assumption underlying this model is that Z and W are conditionally independent given X, which is in agreement with the structure depicted in Figure 2.3 and is necessary for all currently existing three-step approaches. What is also required is that one specifies the distributional form of $P(Z = z|X = t)$. The parameters of the model in Equation 2.11 can be estimated by maximizing the following log likelihood function:

$$\log L_{ML} = \sum_{i=1}^{N} \log \sum_t P(X = t)P(Z = z|X = t)P(W = s|X = t). \tag{2.12}$$

This can be achieved with any software for LCA that can accommodate parameters fixed to some specific values. We fix $P(W = s|X = t)$ to the estimates from step 2.

The possibility of using Z variables of different scale types requires that one should be able to specify an appropriate distribution for Z. Logical choices are a normal distribution for continuous Z, a multinomial distribution for nominal or ordinal Z, a Poisson distribution for count Z, and so forth.

2.4.2 The Bolck-Croon-Hagenaars (BCH) approach

The ML correction method described above uses the classification errors from step two directly in a latent class model for W and Z. In contrast, the solution developed by Bolck et al. (2004) for categorical external predictor variables - which we refer to as the BCH approach - involves re-expressing the relationship described in Equation 2.10 as follows:

$$P(X = t, Z = z) = \sum_s P(W = s, Z = z)d^*_{st}, \tag{2.13}$$

where d^*_{st} represents an element of the inverted $T-$by$-T$ matrix D with elements $P(W = s|X = t)$.[5]

In other words, if we weight the $W - Z$ distribution by the inverse of the classification errors we obtain the distribution we are interested in. Bolck et al. (2004) proposed using

[5]Using matrix algebra, we can write Equation 2.10 as $\mathbf{E} = \mathbf{AD}$, where $\mathbf{E}$ contains the $P(W = s, Z = z)$, $\mathbf{A}$ the $P(X = t, Z = z)$, and $\mathbf{D}$ the $P(W = s|X = t)$. Standard matrix operation yields $\mathbf{A} = \mathbf{ED^{-1}}$ which is what is expressed in Equation 2.13.

this relation which applies at the population level to reweight the data on W and Z (the frequency table with observed counts n_{zs}). As shown by Vermunt (2010), their approach involves maximizing the following pseudo (or weighted) log-likelihood function:

$$\begin{aligned}\log L_{BCH} &= \sum_z \sum_s n_{zs} \sum_{t=1}^{T} d^*_{st} \log P(X=t, Z=z) \\ &= \sum_z \sum_{t=1}^{T} n^*_{zt} \log P(X=t, Z=z), \end{aligned} \tag{2.14}$$

where the $n^*_{zt} = \sum_s n_{zs} d^*_{st}$ are the reweighted frequencies used to estimate the relationship between X and Z.

2.4.3 The modified BCH approach

Vermunt (2010) highlighted three shortcomings of the BCH method: only categorical predictors can be used, standard errors are underestimated, and the method needs a tedious data preparation stage which has to be repeated for each external variable. To solve these issues, the author proposed a modification to the BCH method consisting in reexpressing the pseudo log-likelihood function in terms of individual observations. That is,

$$\begin{aligned}\log L_{BCH} &= \sum_{i=1}^{N} \sum_{s=1}^{T} w_{is} \sum_{t=1}^{T} d^*_{st} \log P(X=t, Z=z) \\ &= \sum_{i=1}^{N} \sum_{t=1}^{T} w^*_{it} \log P(X=t, Z=z), \end{aligned} \tag{2.15}$$

where w_{is} is a class assignment weight and $w^*_{it} = \sum_{s=1}^{T} w_{is} d^*_{st}$. Note that the standard three-step procedure involves using the non-reweighted w_{is} in the third step. In order to apply this modified BCH method, an expanded data file has to be created containing T records for each subject with X values $t = 1, 2, 3...T$ and weights w^*_{it}. This weighted data set can be analyzed with standard methods. While Equation 2.15 shows how to estimate parameters of the joint distribution of X and Z, it can be modified for the estimation of the conditional distribution of Z given X as follows:

$$\begin{aligned}\log L_{BCH} &= \sum_{i=1}^{N} \sum_{t=1}^{T} w^*_{it} \log P(X=t) P(Z=z|X=t) \\ = \sum_{i=1}^{N} \sum_{t=1}^{T} w^*_{it} \log P(X=t) &+ \sum_{i=1}^{N} \sum_{t=1}^{T} w^*_{it} \log P(Z=z|X=t). \end{aligned} \tag{2.16}$$

Because the first term does not contain parameters of interest it can be ignored and we can just maximize a pseudo log-likelihood function based on the second term. Note that this formulation makes it possible to apply the BCH method to external variables of any scale type, thus also with continuous and ordinal Z variables. By applying a robust or sandwich variance estimator, one can prevent that standard errors (SEs) are underestimated as is the case with the original BCH approach. The robust variance-covariance matrix of the parameters is the inverse of the matrix obtained by "sandwiching" the Hessian by the average outer product of gradients for the independent observations (Skinner, Holth and Smith 1989).

2.4.4 ML adjustment with multiple latent variables

For the clarity of exposition, so far we have focused on the situation in which the step three latent class model of interest contains only one latent variable. However, both the ML and BCH method can easily be extended to be applicable with multiple latent variables. We will illustrate this for the somewhat simpler ML approach.

Suppose one is interested in the association between latent variables X_1 and X_2. A stepwise modeling approach implies that one performs a separate LCA for each of these two latent variables and obtains class assignments W_1 and W_2. Implicitly, this means that an additional assumption is made, namely that the indicators used in the model for X_1 are independent of X_2 conditionally on X_1 and vice versa. Given these assumptions are met, it is no problem to estimate the measurement models separately. The relationship between the joint distribution of the assigned class memberships and the true class memberships can be expressed similar to Equation 2.10 as follows:

$$\begin{aligned} P(W_1 = s_1, W_2 = s_2) = \sum_{t1} \sum_{t2} P(X_1 = t_1, X_2 = t_2) \\ P(W_1 = s_1 | X_1 = t_1) P(W_2 = s_2 | X_2 = t_2). \end{aligned} \tag{2.17}$$

This is a latent class model that can be estimated using LCA packages that support the use of multiple latent variables (here X_1 and X_2) and fixed value parameters [here $P(W_1 = s_1|X_1 = t_1)$ and $P(W_2 = s_2|X_2 = t_2)$]. As shown for the $X - Z$ association, rather than modeling the joint distribution of X_1 and X_2, it is also possible to model the conditional distribution $P(X_2 = t_2|X_1 = t_1)$, also when observed predictors are included in the model - $P(X_2 = t_2|X_1 = t_1, Z = z)$. Moreover, extension to more than two latent variables is straightforward. We illustrate the use of this method with our second real data example.

The generalized correction methods introduced above will be tested in the following with a simulation study and illustrated with two real data examples. For ease of readability, the simulation study focuses on the situation with one independent latent variable and one dependent variable. The extension to more complex models is shown using the examples. In order to show the ease of use and applicability with the real data example, the syntax used in Latent Gold (Vermunt and Magidson, 2013) will be included as well. Since Vermunt (2010) showed that the SE's are underestimated using the original BCH method, here we will use only the modified BCH method with robust standard errors.

2.5 Simulation study

2.5.1 Design

A simulation study was conducted in order to check the quality of the proposed adjusted three-step LCA methods in situations in which the latent variable is treated as a predictor of one or more external variables (distal outcomes). In the simulation study, the BCH and ML correction methods were compared with the one-step and the standard three-step approach. A method can be considered to perform well when the parameter estimates are unbiased and their variation is small, and in general the estimates are accurate. In the simulation study we will manipulate two key factors: the separation between classes (which as explained earlier is strongly related to the size of the classification error)[6] and the sample size, which both have been found to affect the performance of the correction methods when the three-step LCA involved prediction of class membership using external variables (Vermunt 2010). Separation between classes is manipulated via the strength of the relationship between the classes and the indicators. Other conditions that could have been varied are number of items, number item categories, and class sizes, but these are all conditions that basically affect the separation between classes. To keep the simulation simple and manageable, we decided to manipulate class separation only via the class-item association.

We tested the performance of the correction methods for three types of distal outcomes; that is, for Z nominal, ordinal, or continuous. Two conditions were used for the strength of the $X - Z$ relationship, corresponding to a weaker and a stronger effect of X on Z. Data were generated from the full (X, Y, Z) model. In the following the population values for all the parts of the model are provided. The population model we used is a three-class model for six dichotomous response variables and a single distal outcome variable. The profile of the classes is as follows: class one is likely to give the high response on all indicators, class two scores high on the first three indicators and low on the last three, and class three is likely to give the low response on all indicators. The separation between classes was manipulated by changing the conditional response probabilities for the indicators. The probability for the likely response was set to .70, .80, and .90, corresponding to a (very) low, middle, and high separation between classes. These settings correspond with entropy based R^2 values of .36, .65, and .90, respectively. In the following we will refer to these conditions as the low, mid, and high separation condition. Sample size is also important because it affects the accuracy of the estimates. The three sample sizes used were 500, 1000, and 10000. Note that a class separation of .36 is in fact extremely low and a sample size of 10000 is rather large. We used three types of outcome variables, a trichotomous nominal, a trichotomous ordinal, and a continuous outcome, which we modeled using a multinomial logit, a cumulative logit, and a linear model, respectively, with the first class and the first category of the outcome variable as the reference category.

For the nominal outcome, the condition with a strong effect of X on Z was obtained by setting the intercepts β_2 and β_3 to -2.08 and the effect parameters to 3.87 (β_{22}),

[6] The separation is measured by the Entropy R^2, which tells how much the prediction of X improved when using the information on $\mathbf{Y}$. If $P(X = t|\mathbf{Y} = \mathbf{y})$ is close to 0 or 1 for most data patterns, the separation between the classes is good, and the classification error is low.

3.17 (β_{23}), 2.08 (β_{32}), and 2.08 (β_{33}), where the first index refers to the distal outcome category and the second to the class. Note that this set up yields some probabilities close to 0, which can cause estimation problems, as we will see in the Results section. For the condition with a weaker effect of X on Z we set both intercepts equal to -1.098, $\beta_{22}$to 2.01, β_{23} to 1.50, β_{32} to 2, and β_{33} to 1.09.

For the ordinal outcome variable, in the high effect condition the thresholds were set to 2.94 (β_2), 1.55 (β_3), and the effect parameters to -1.55 (β_2) and -4.33 (β_2), for class 2 and 3 respectively. This setup also yields some probabilities close to 0. In the low effect condition, the thresholds were set to 2.74 (β_2) and 1.82 (β_3) and the effect parameters to -1.23 (β_2) and -3.01(β_3).

For the continuous outcome variable, in the strong effect condition, we set the class specific means to -1, 0 and 1 (corresponding with an intercept of -1 and slopes of 1 and 2), and the error variance to 1. In the weak effect condition, we set the class specific means equal to -0.2, 0, 0.2, and kept the same error variance.

For the simulation study and the real data application, two computer programs were used: Latent GOLD (Vermunt and Magidson 2013) and R (Venables, Smith, the R Core Team, 2013). In Latent GOLD we simulated the data, set up the measurement model, saved the scores on the posterior class assignment, and run all the correction methods with both modal and proportional assignment. We used R to construct the $\mathbf{D}$ matrix and compute its inverse, and to create the expanded data matrix containing the relevant weights. The $\mathbf{D}$ matrix was computed using Equation 2.6; that is, using the empirical distribution of the responses. For each of the 54 conditions, which were obtained by crossing the 3 separation, 3 sample size, 3 types of external variable, and 2 effect size conditions, we used 500 replications.

2.5.2 Results

The results are presented both averaged across conditions, and separately for some of the conditions. We pay attention to parameter bias (measured by comparing the average estimated value with the true values) and efficiency (measured by the standard deviation across replications), and to the bias in the estimated standard errors (measured by comparing the average estimated standard error with the standard deviation across replications).

Before looking at these figures, we would like to present an important unanticipated result for the BCH method when applied with a nominal or an ordinal outcome variable Z. Some of the replications turned out to contain negative cell frequencies in the adjusted $X - Z$ frequency table, in which case the corresponding multinomial distribution is not defined. This happened mainly in the least favorable condition coupling a low class separation (large classification errors) with a small sample size (large sampling fluctuation). The possibility of such a failure of the BCH method is an important new result because it was not reported by Bolck et al. (2004) or Vermunt (2010). While an ad hoc solution could be to fix the probabilities corresponding to negative counts to zero, we decided to exclude replications with negative frequencies from the results reported below. In the replication samples where the BCH method gave negative frequencies, the three-step ML method gave logit coefficients going to plus or minus infinity, corresponding to boundary

Table 2.1: Number of Excluded Replications for the Nominal and Ordinal Outcome Variable due to Negative Frequencies or Boundary Solutions

Sample size	Separation level	Correction methods	One-step ML
	Nominal - strong X-Z effect		
500	Low	63	200
1000	Low	59	59
500	Mid	4	1
1000	Mid	1	0
	Nominal - weak X-Z effect		
500	Low	9	46
1000	Low	5	4
	Ordinal - strong X-Z effect		
500	Low	20	28
1000	Low	18	0

solutions. Boundary solutions also occurred with the one-step ML method in the low separation and low sample size conditions. The replications with negative frequencies and boundary solutions were excluded from further analysis. Table 2.1 provides information on the number of excluded replications per condition.

Table 2.2 presents the results averaged over all sample sizes and separation levels for one parameter per outcome variable. It reports the average estimate, average SE, and SD of estimates for each method. As can be seen, the proportional standard method has the largest bias. When averaged across conditions, we can see that the correction methods still slightly underestimate the parameters. The bias is less than 5% for the continuous and ordinal outcome variable, and close to 10% for the nominal outcome variable. As shown below, bias varies strongly across separation and sample size conditions (is larger with low separation and small sample size, and absent with higher separation and large sample size). As expected, when estimating a correctly specified model, the one-step approach yields a good approximation of the parameter of interest (bias less than 5%). It should be mentioned that with the exception of the low sample size and low separation between classes conditions the correction methods perform well, having bias less than 5% for all outcome variables as well.

As can be seen from the standard deviations across replications (SD's), the correction methods perform similar in terms of efficiency with each other and the one-step ML method. Comparison of the average estimated SE across replications with the SD of the parameter estimate across replications shows that the correction methods slightly underestimate the SE, with the exception of the proportional ML method, which overestimates the SE for the nominal outcome variable. Overall, the difference between the SE's and SD's is smallest for the proportional ML method, except for the nominal outcome variable.

When we look at the parameter estimates separately in each of the investigated conditions, we see large differences between conditions. As seen in Tables 2.3 and 2.4, the one-step ML method obtains estimates close to the true values, with the exception of the combination of small sample size and low separation between classes, where it tends to

Table 2.2: Average Estimate of One Selected β Parameter, and its Average Estimated SE and SD across Replications Aggregated over the Nine Separation and Sample Size Conditions for all Three types of Outcome Variables (for strong and weak X-Z association)

	Nominal			Ordinal			Continuous		
Method	Estimate	SE	SD	Estimate	SE	SD	Estimate	SE	SD
	β_{23} = 3.17			β_2 = -1.56			β_1 = 1.00		
One-step ML	3.22	0.55	0.50	-1.58	0.27	0.27	1.00	0.07	0.07
Modal standard	2.06	0.22	0.23	-1.18	0.15	0.20	0.80	0.06	0.08
Proportional standard	1.73	0.21	0.18	-1.07	0.15	0.16	0.72	0.06	0.07
Modal BCH	2.97	0.50	0.51	-1.52	0.26	0.33	0.97	0.07	0.09
Proportional BCH	2.98	0.56	0.48	-1.55	0.25	0.32	0.97	0.07	0.10
Modal ML	2.97	0.50	0.51	-1.52	0.25	0.31	0.97	0.07	0.09
Proportional ML	2.98	0.83	0.51	-1.53	0.30	0.30	0.97	0.07	0.08
	β_{23} = 1.50			β_2 = -1.23			β_1 = 0.20		
One-step ML	1.53	0.40	0.35	-1.26	0.25	0.26	0.20	0.07	0.06
Modal standard	1.05	0.21	0.22	-0.97	0.15	0.17	0.16	0.05	0.05
Proportional standard	0.90	0.19	0.15	-0.88	0.14	0.14	0.14	0.04	0.05
Modal BCH	1.42	0.31	0.32	-1.22	0.23	0.26	0.19	0.06	0.06
Proportional BCH	1.42	0.29	0.30	-1.24	0.22	0.26	0.20	0.06	0.06
Modal ML	1.42	0.31	0.32	-1.22	0.22	0.26	0.19	0.06	0.06
Proportional ML	1.42	0.41	0.30	-1.23	0.28	0.26	0.20	0.07	0.06

overestimate the parameter. For all outcome variables, the correction methods perform poorly in the low separation and small sample size conditions, a result that is similar to the one reported by Vermunt (2010). Note that this applies to each of the three types of response variables and both for a strong and a weak $X - Z$ association. The reason for this bad performance with low separation and small sample size is that in this situation the differences between classes are overestimated in the first step yielding an underestimate of (a too optimistic) classification error, and as a consequence a too moderate adjustment by the BCH and ML correction methods. In the middle and high separation conditions, the correction methods perform well. While in the high separation conditions the performance of the correction methods using modal versus proportional assignment did not differ, in the lower separation condition this is not the case. With middle separation and especially with low separation between classes, the estimates obtained with the proportional assignment approximated better the true values than the ones obtained using modal assignment for all three types of outcome variables.

Table 2.5 reports the average SE and SD across replications for one selected parameter (from the condition with a nominal Z variable weakly related to the classes) for the nine sample size and class separation combinations. As we can see, in the conditions with a low separation and a smaller sample size the proportional ML and one-step ML method tend to overestimate parameter uncertainty (SE is higher than SD). The other correction methods slightly underestimate the SE's in all nine conditions. With regard to efficiency the correction methods perform similar to the one-step ML method, with the exception of the combination of small sample size coupled with low separation, for which

Table 2.3: Average Estimate of Selected β Parameter Separately for each of the Nine Separation and Sample Size Conditions for all Three types of Outcome Variables for strong X-Z association

Separation level	Low			Medium			High		
Sample size	500	1000	10000	500	1000	10000	500	1000	10000
Method	Nominal Z: $\beta_{23} = 3.17$								
One-step ML	3.28	3.24	3.11	3.36	3.22	3.17	3.20	3.18	3.16
Modal standard	1.14	1.20	1.34	2.07	2.11	2.13	2.85	2.83	2.82
Proportional standard	0.90	0.87	0.85	1.69	1.69	1.68	2.63	2.61	2.60
Modal BCH & ML	2.03	2.40	3.16	3.17	3.24	3.17	3.21	3.18	3.15
Proportional BCH & ML	2.13	2.58	3.11	3.11	3.15	3.16	3.18	3.16	3.15
	Ordinal Z: β_2 = -1.56								
One-step ML	-1.64	-1.61	-1.56	-1.60	-1.57	-1.56	-1.59	-1.56	-1.56
Modal standard	-0.83	-0.84	-0.84	-1.22	-1.21	-1.22	-1.51	-1.49	-1.48
Proportional standard	-0.67	-0.65	-0.65	-1.11	-1.09	-1.09	-1.46	-1.44	-1.44
Modal BCH	-1.40	-1.41	-1.54	-1.55	-1.52	-1.55	-1.58	-1.56	-1.56
Proportional BCH	-1.49	-1.50	-1.56	-1.57	-1.55	-1.56	-1.58	-1.56	-1.56
Modal ML	-1.39	-1.42	-1.54	-1.56	-1.52	-1.55	-1.58	-1.56	-1.56
Proportional ML	-1.43	-1.42	-1.56	-1.57	-1.56	-1.55	-1.58	-1.56	-1.56
	Continuous Z: $\beta_1 = 1.00$								
One-step ML	1.00	1.00	0.99	0.99	1.00	1.00	1.03	1.00	1.00
Modal standard	0.57	0.59	0.60	0.81	0.83	0.84	0.96	0.96	0.96
Proportional standard	0.47	0.47	0.45	0.74	0.75	0.75	0.94	0.94	0.94
Modal BCH	0.85	0.91	0.98	0.97	0.99	1.00	1.00	1.00	1.00
Proportional BCH	0.88	0.94	0.98	0.97	1.00	1.00	1.00	1.00	1.00
Modal ML	0.85	0.91	0.99	0.97	1.00	1.00	1.00	1.00	1.00
Proportional ML	0.87	0.93	0.99	0.98	1.00	1.00	1.00	1.00	1.00

the correction methods are more efficient. Comparison of these results to those for other outcome variables and effect sizes showed that the SE bias of the correction methods is slightly larger in the strong effect condition for nominal and ordinal outcomes, and smaller for continuous outcomes irrespective of the effect size.

2.6 Two empirical examples

2.6.1 Example 1: Psychological contract types

To illustrate the working of the correction methods, we analyzed data from the Dutch and Belgian sample of the Psychological Contracts across Employment Situation (PSYCONES) project (European Commission, 2006). We used the same questionnaire items as De Cuyper, Rigotti, De Witte and Mohr (2008) who performed a LCA to build a typology for psychological contracts between employers and employees. Out of the 8 dichotomous indicators, 4 refer to employees obligations (whether a promise was made or not) and 4 to employers obligations, where each set of 4 items contained 2 items for relational and 2 for transactional obligations. Examples of the wording of items are: 'This organization promised me a reasonably secure job' and 'This organization promised me a good pay for

Table 2.4: Average Estimate of Selected β Parameter Separately for each of the Nine Separation and Sample Size Conditions for all Three types of Outcome Variables for weak X-Z association

Separation level	Low			Medium			High		
Sample size	500	1000	10000	500	1000	10000	500	1000	10000
Method	Nominal Z: β_{23} =1.50								
One-step ML	1.53	1.61	1.51	1.57	1.53	1.51	1.51	1.51	1.50
Modal standard	0.61	0.66	0.72	1.10	1.10	1.10	1.39	1.40	1.39
Proportional standard	0.49	0.49	0.47	0.93	0.91	0.90	1.30	1.31	1.30
Modal BCH & ML	1.02	1.22	1.46	1.50	1.52	1.51	1.50	1.51	1.50
Proportional BCH & ML	1.06	1.29	1.44	1.50	1.50	1.51	1.50	1.51	1.50
	Ordinal Z: β_2 = -1.23								
One-step ML	-1.33	-1.29	-1.25	-1.25	-1.25	-1.23	-1.25	-1.27	-1.24
Modal standard	-0.72	-0.72	-0.72	-0.99	-1.01	-0.99	-1.20	-1.22	-1.19
Proportional standard	-0.59	-0.57	-0.55	-0.91	-0.91	-0.90	-1.17	-1.19	-1.16
Modal BCH	-1.14	-1.18	-1.23	-1.21	-1.24	-1.22	-1.25	-1.27	-1.24
Proportional BCH	-1.22	-1.23	-1.25	-1.23	-1.24	-1.23	-1.25	-1.27	-1.24
Modal ML	-1.14	-1.18	-1.24	-1.22	-1.25	-1.23	-1.25	-1.27	-1.24
Proportional ML	-1.21	-1.22	-1.24	-1.23	-1.24	-1.23	-1.22	-1.27	-1.24
	Continuous Z: β_1= 0.20								
One-step ML	0.21	0.20	0.20	0.21	0.20	0.20	0.20	0.20	0.20
Modal standard	0.12	0.12	0.12	0.17	0.16	0.17	0.19	0.19	0.19
Proportional standard	0.10	0.09	0.09	0.16	0.15	0.15	0.18	0.18	0.19
Modal BCH	0.18	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
Proportional BCH	0.19	0.20	0.20	0.20	0.19	0.20	0.20	0.20	0.20
Modal ML	0.18	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
Proportional ML	0.19	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20

the work I do'. The sample consisted of 1353 respondents. The distal outcome variable Z was the perceived job insecurity measured using a scale developed by De Witte (2000). This scale consists of 4 items with 5 categories and had a Cronbach's alpha value of .88.

In the first step, we fitted the measurement model using the eight indicator variables. Based on the BIC values and the bivariate residuals between the items, it was concluded that a four-class model fitted the data well. Table 2.6 presents the parameter estimates for this four-class model. Class 1 (9 % of respondents) is characterized by mutual low obligations. Class 2 (10%) represents employee under obligation: these respondents are likely to perceive employers obligations as given, and have a lower probability of perceiving own obligations as promised. Class 3 (29%) represent employees who themselves made promises to the organization, but received less: the over obligation class. Class 4 (52%) scores high on all items, representing mutual high obligations.

After identifying the classes, the posterior class membership probabilities were saved, and the $\mathbf{D}$ matrix with elements $P(W = s|X = t)$ and its inverse were calculated (note the calculations of the weighting happens behind the scenes for the version 5.00 of Latent GOLD). The one-step and the corrected and uncorrected three-step methods were used to analyze the relationship between class membership and perceived job insecurity, where

Table 2.5: Average Estimated SE and SD across replications for all nine conditions separately for One Parameter for nominal outcome variable Z ($\beta_{23} = 1.50$) obtained using the One-step ML and the step three correction methods

Sample size	500		1000		10000	
Method	SD	SE	SD	SE	SD	SE
	Low separation					
One-step ML	0.88	1.28	0.64	0.74	0.09	0.09
Modal BCH	0.66	0.65	0.53	0.51	0.16	0.15
Proportional BCH	0.61	0.58	0.48	0.46	0.15	0.14
Modal ML	0.66	0.64	0.53	0.51	0.16	0.15
Proportional ML	0.61	0.90	0.48	0.79	0.15	0.27
	Mid separation					
One-step ML	0.46	0.43	0.33	0.30	0.09	0.09
Modal BCH	0.47	0.45	0.34	0.31	0.10	0.10
Proportional BCH	0.44	0.42	0.32	0.29	0.09	0.09
Modal ML	0.47	0.45	0.34	0.31	0.10	0.10
Proportional ML	0.44	0.55	0.32	0.39	0.09	0.12
	High separation					
One-step ML	0.34	0.33	0.22	0.23	0.07	0.07
Modal BCH	0.34	0.33	0.22	0.23	0.07	0.07
Proportional BCH	0.34	0.33	0.22	0.23	0.07	0.07
Modal ML	0.34	0.33	0.22	0.23	0.07	0.07
Proportional ML	0.34	0.35	0.22	0.25	0.07	0.08

the latter is treated as a continuous variable with a constant error variance; that is, in the three step approaches we used a linear regression to regress job insecurity on class membership, and in the one step method we used job insecurity as a continuous indicator variable. This is the relevant part of the Latent GOLD 5.00 syntax used for three-step ML with modal assignment:

```
"step3 modal ML
variables:
latent cluster nominal posterior = ( cluster1 cluster2 cluster3 cluster4 );
dependent insecurity continuous;
equations
Insecurity <- 1 + cluster;
Insecurity"
```

The estimated effect sizes of psychological contract type on job insecurity (and their SE's) and the value of the Wald test for the overall effect (and its p value) are reported in Table 2.7. As we can see in the table, the job insecurity of the employee under obligation (class 2) and mutual high obligations group (class 3) is lower than for those in the mutual low obligation group (class 1). The job insecurity of the employee over obligation group is similar to that of the mutual low obligation group. Comparing the effect parameters obtained by the different methods, we can see that the standard three-step procedures yield estimates that are far away from the ones of the other methods, while all the other methods yield similar estimates. The correction methods have slightly smaller parameter

Table 2.6: Class Proportions and Class-Specific Probabilities of a Positive Response for the Four-Class Model Estimated for the PSYCONES Data

	Class 1 Mutual low	Class 2 Under-obligation	Class 3 Over-obligation	Class 4 Mutual high
Class proportion	.09	.10	.29	.52
Employers' obligations				
Secure job	.21	.87	.36	.90
Advancement	.18	.85	.30	.90
Good pay	.26	.75	.28	.87
Safe work environment	.29	.73	.55	.97
Employees' obligations				
Loyalty	.08	.36	.73	.98
Volunteer	.17	.37	.83	.98
On time	.18	.39	.96	.98
Good performance	.28	.77	.97	.99

Table 2.7: Effect of Class Membership on Job Insecurity, SE's, Multivariate Wald Test for the Effect, and its Significance Obtained with the Seven Different Methods, using Dummy Coding with First Class as Reference Category.

Method	Class 2 (SE)	Class 3 (SE)	Class 4 (SE)	Wald (DF)	p
One-step ML	-0.59 (0.16)	0.09 (0.13)	-0.45 (0.11)	63.67 (3)	<.001
Modal standard	-0.42 (0.13)	0.01 (0.10)	-0.36 (0.10)	47.21 (3)	<.001
Proportional standard	-0.34 (0.12)	0.01 (0.10)	-0.34 (0.10)	39.77 (3)	<.001
Modal BCH	-0.53 (0.17)	0.03 (0.13)	-0.42 (0.11)	44.58 (3)	<.001
Proportional BCH	-0.51 (0.16)	0.06 (0.12)	-0.42 (0.11)	53.43 (3)	<.001
Modal ML	-0.54 (0.16)	0.04 (0.13)	-0.43 (0.11)	48.16 (3)	<.001
Proportional ML	-0.54 (0.19)	0.08 (0.14)	-0.43 (0.12)	41.19 (3)	<.001

values than the one-step method, where the three-step ML methods are closer to the one-step method than the BCH methods. Similarly to the results of the simulation study, the SE's obtained using the proportional ML method are slightly larger than the ones obtained using the other correction methods and the one-step ML. The SE's obtained by the other correction methods are similar to the ones obtained using one-step ML method. Looking at the Wald tests of the correction methods, it can be seen that the Wald value for the proportional ML method is the lowest, meaning that this method is the most conservative.

2.6.2 Example 2: Political ideology

In many research situations, it is of interest to predict a latent outcome variable from other latent variables. We will illustrate how the ML three-step method can be used for this purpose with data from the Dutch sample of the 1981 European Value Survey (GESIS-Variable Reports No. 2011/06). More specifically, we investigate how religiosity

Table 2.8: Class Proportions and Class-Specific Probabilities of Religiosity for the Three-Class Model Estimated for the 1981 wave of the EVS data

		Class 1 Non-religious	Class 2 Middle	Class 3 Religious
Class proportion		.34	.33	.33
Religiosity	No	.95	.06	.01
	Yes	.05	.94	.99
Personal God	No	.99	.80	.11
	Yes	.01	.20	.89
Traditionalism	Nontraditional	.84	.15	.01
	Intermediate	.15	.65	.06
	Traditional	.01	.20	.93
Religious org. membership	No	.96	.66	.25
	Yes	.04	.34	.75
Denomination	Yes	.08	.80	.99
	No	.92	.20	.01
Prayer	Yes	.32	.66	.96
	No	.68	.33	.04

affects political ideology while controlling for social status. Social status is an observed variable with four ordinal categories: professional/managerial, semi-skilled, unskilled or unemployed or pensioner (1); skilled manual workers (2); sales, clerical and other non-manual (3); above average life style (4). We used social status as a numeric covariate in the analysis. Similarly to Hagenaars and Halman (1989) work on the same dataset, we modeled political ideology and religiosity as categorical latent variables measured using multiple indicators.

Religiosity was measured with six indicators, among which praying, belonging to a church, and belonging to a denomination (see Table 2.8). We selected the three-class model based on the BIC and the goodness-of fit (L2=83.68, df=74, p=0.21). Class separation is good (entropy R^2=.85).Table 2.8 shows the class solution. Group one, (34%) are the 'non religious' scoring low on all items, while group 2 the 'middle' (33%) has mixed scores, and the last group the 'religious' (33%) score high on all items.

Political ideology was measured with six indicators, among which party closeness and left-right orientation (see Table 2.9). We fitted LC models with different numbers of classes and selected the three-class model based on the lowest BIC and a nonsignificant goodness-of-fit statistic (L2 =79.38, df=74, p=0.31). Class separation is moderate (entropy R^2=.68). Group one can be characterized as 'left wing' (27%), group two as 'middle/indifferent' (37%), and the last group as 'right wing' (35%).

Note that the two measurement models are estimated separately. Each yields a set of modal class assignments and an estimate of the $\mathbf{D}$ matrix with the conditional probability of being assigned to a class conditional on the true class membership. The class assignments and the two $\mathbf{D}$ matrices can be used to set up the model in which (latent)

Table 2.9: Class Proportions and Class-Specific Probabilities of Political mentality for the Three-Class Model Estimated for the 1981 wave of the EVS data

Class:		Class 1 Left	Class 2 Middle/Indifferent	Class 3 Right
Class proportion		.27	.37	.35
Left/right	Left	.89	.25	.02
	Middle	.10	.53	.27
	Right	.01	.22	.71
Political interest	No	.28	.77	.39
	Yes	.72	.23	.61
Trust in parliament	No	.60	.68	.26
	Yes	.40	.32	.74
Societal change	No	.17	.20	.41
	Yes	.83	.80	.59
Equality vs freedom	Equality	.51	.60	.76
	Freedom	.49	.40	.24
Party closeness	Yes	.92	.10	.83
	No	.08	.90	.17

political ideology is predicted from social status and (latent) religiosity. The syntax for the specified model using the three-step ML method is provided in Appendix 1 with the automated Latent GOLD syntax, and the version with the user defined $\mathbf{D}$ matrices as well. Table 2.10 reports the estimates for the effects of religiosity and social class on political ideology obtained with the one-step, the adjusted three-step, and the standard unadjusted three-step approach. As can be seen, the results obtained with all three methods point toward the same tendencies. While the estimates from the one-step and the corrected three-step method are rather similar, the uncorrected three-step approach yields smaller effect sizes.

Based on the estimated multinomial logit coefficients one can conclude that controlling for social class, the more religious a person, the more likely it is that (s)he is political right or middle/indifferent rather than left. Moreover, the higher the social class the more likely to be rightwing rather than leftwing, while there is no significant effect of social class on the middle-left contrast.

2.7 Discussion

We proposed a generalization of existing correction methods for the attenuation problem appearing in three-step LCA with external variables. We showed how two existing correction methods for latent class models with covariates can be generalized to a broader range of situations; that is, to formulate models for the joint probability of class membership and external variables. The correction methods can therefore now be applied in any situation where we wish to relate scores on class membership with external variables, irrespective

Table 2.10: Multinomial Logit Coefficients from the Regression of Religiosity on Social Class on Political Ideology, SE's, Multivariate Wald Tests, obtained with Three Different Methods using Dummy Coding with the First Class as Reference Category for Religiosity and Political Ideology

	Political= Middle (SE)	Political= Right (SE)	Wald(DF)
One step ML			
Religiosity =Middle	0.67(0.35)	0.80(0.41)	52.80(4)
Religiosity =Religious	1.23(0.57)	3.09(0.50)	
Social Class	-0.27(0.47)	1.39(0.39)	18.33(2)
Modal ML			
Religiosity =Middle	0.48(0.34)	0.72(0.45)	39.85(4)
Religiosity =Religious	1.10(0.46)	2.76(0.48)	
Social class	-0.15(0.43)	1.33(0.41)	14.83(2)
Uncorrected Modal			
Religiosity =Middle	0.43(0.27)	0.65(0.31)	47.94(4)
Religiosity =Religious	0.99(0.31)	2.10(0.33)	
Social Class	0.04(0.33)	1.03(0.32)	16.74(2)

of the hypothesized causal order. Though we focused mainly on the situation in which class membership is a predictor of a continuous, ordinal, or nominal outcome variable, the correction methods can be applied in relation with distal outcome variables having almost any of the distributional forms from the exponential family. We also showed how the ML correction method can be extended to models with more than one latent variable.

The performance of the correction methods was tested by a simulation study and illustrated with two real data examples. The results of the simulation study show, similarly to previously reported results, that using the uncorrected three-step approach leads to seriously biased parameter estimates of the association of class membership with external variables. Although the direction of the effects is correct, the effect sizes are very much attenuated. As such it is recommended to use one of the correction methods when deciding to use the three-step approach. All correction methods we tested perform well; both their estimates and SE's can be trusted, with the exception of the situations where the class separation of the measurement model is very low, in which situation they underestimate the parameter estimates and SE's. The most efficient correction method is the proportional ML method. In general, the results obtained with proportional assignment are better than those obtained using modal assignment. A non-anticipated result is that for nominal outcomes the BCH method may fail because of the occurrence of negative cell frequencies, a problem that is much more likely to occur with a (very) low separation between classes. Therefore, the use of the one-step or three-step ML methods is recommended in these situations.

One of the limitations of the current study is that it examined the behavior of the correction methods only for the situation in which model assumptions hold; that is, we did not look at situations in which distributional assumptions about the external variables and/or conditional independence assumptions are violated. Our expectation is that the adjusted three-step methods may perform better than the one-step method under misspecification, which is one of the issues we will focus on in future research. Another limitation is that we focused mainly on parameter bias and less on hypothesis testing. This means that no statements can be made about issues such as amount of power decrease of statistical tests such as the Wald test resulting from using the proposed correction methods. This is another topic for future research.

Chapter 3

Stepwise LCA: Standard errors for correct inference

Abstract

Latent class analysis is used in the political science literature in both substantive applications and as a tool to estimate measurement error. Many studies in the social and political sciences relate estimated class assignments from a latent class model to external variables. Though common, such a "three-step" procedure effectively ignores classification error in the class assignments. Vermunt (2010) showed that this leads to inconsistent parameter estimates and proposed a correction. Although this correction for bias is now implemented in standard software, inconsistency is not the only consequence of classification error. We demonstrate that the correction method introduces an additional source of variance in the estimates, so that standard errors and confidence intervals are overly optimistic when not taking this into account. We derive the asymptotic variance of the third-step estimates of interest, as well as several candidate corrected sample estimators of the standard errors. These corrected standard error estimators are evaluated using a Monte Carlo study and we provide practical advice to researchers as to which should be used so that valid inferences can be obtained when relating estimated class membership to external variables.

This chapter is published as Bakk, Z., Oberski, D.L. & Vermunt, J. K. (2014). Relating latent class membership to continuous distal outcomes: improving the LTB approach and a modified three-step implementation. *Political Analysis, vol 22, pp. 520-540*

3.1 Introduction

Latent class analysis (LCA) is a tool used to classify objects for further analysis (Ahlquist & Breunig, 2012), with a wide range of applications in political science. For example, McCutcheon (1985) examined the effect of education and age cohort on Americans' tolerance for nonconformity as obtained from a latent class model; Mustillo (2009, Table 4) provided a hard partitioning of new political parties in volatile party systems. Furthermore Grimmer and Stewart (2013) discuss latent class analysis as an unsupervised machine learning method for political texts such as debates, legislation, news reports, and party manifestos; and Grimmer (2013) related latent classes obtained from US Senators' press releases to their publicly expressed priorities. Further applications of LCA in political science include Feick (1989); Sniderman, Tetlock, Glaser, Green, and Hout (1989);Breen (2000), Hill and Kriesi (2001); Blaydes and Linzer (2008); Linzer (2011); Glasgow, Golder, and Golder (2012); Ristei Gugiu and Centellas (2013) and Beissinger (2013). While most applications we refer to are substantive, LCA, and in general latent variable models can be used in a more instrumental manner as well, as a tool to estimate measurement error in observed variables (Fuller, 1987; Alwin, 2007; Fornell & Larcker, 1981; Rabe-Hesketh, Skrondal, & Pickles, 2001; Oberski & Satorra, 2013).

As the above examples already suggest in most applications the interest lies not only in creating a latent classification, but also in relating this classification to external variables of interest. Usually this is done using a three-step procedure, even though a simultaneous estimation procedure is also available (Dayton & Macready, 1988; Hagenaars, 1990, 1993; Bandeen-Roche et al., 1997; van der Heijden, Dessens, & Bockenholt, 1996). The three-step approach proceeds as follows: in the first step, the latent class model is estimated; secondly units are classified into classes using some assignment mechanism based on the first step; and thirdly the newly created observed variable is related to external variables using standard methods such as (logistic) regression. Note that in the second step a classification error is introduced because the true class membership is unknown, unless there is perfect classification, and this error leads to biased parameter estimates in the third step (Vermunt, 2010; Bolck, Croon, & Hagenaars, 2004).

Bias notwithstanding the three-step approach is still very popular in applied social and bio-medical research (Olino et al., 2011; McCutcheon, 1985; R. M. Clark & Besterfield-Sacre, 2009; Marsh, Ludtke, Trautwein, & Morin, 2009; Loken, 2004; Chan & Goldthorpe, 2007). This popularity can be explained among other factors by the intuitive nature of the approach: researchers prefer to first establish a measurement model or a construct, and later regress the construct on potential predictors (Vermunt, 2010). The stepwise approach is preferred even more in situations where the classification needs to be related to dependent variables (distal outcomes). The reason for this preference is that if the dependent variables are added in a single step these variables would define the classification, whereas the intent is to explain them by the classification (Bakk, Tekle, & Vermunt, 2013; Lanza, Tan, & Bray, 2013), thus an unintended circularity would be created. In many situations the different steps are performed by different researchers, at different points in time. The stepwise approach can also be used in situations where the simultaneous estimation would be impossible, for instance when information about the classification error comes from a different sample.

Inspired by the widespread use of the three-step approach, Vermunt (2010) provided an improved three-step procedure in which the third step is amended by correcting for classification errors, thus removing the parameter bias. Bakk et al. (2013) and Asparouhov and Muthén (2014) tested Vermunt's approach via simulation studies by using models with distal outcome variables and latent transition analysis respectively, showing that in all these situations the bias-adjusted three-step approach performs well with regard to parameter bias reduction. Furthermore Feingold, Tiberio, and Capaldi (2013) applied the corrected three-step approach to substance abuse data, and implementations of the method are available in standard latent class software Mplus Version 7.1 (Asparouhov & Muthén, 2014) and Latent GOLD Version 5.00 (Vermunt & Magidson, 2013).

As such the improved three-step procedures of Vermunt (2010) are easy to use due to their availability in mainstream software. However, as we show in this article, even after correcting for parameter bias an additional source of error remains, namely the three-step procedure causes additional variance in the estimates that should be accounted for. Depending on the assignment method used in step two, standard errors will be over- or underestimated (Vermunt, 2010; Bakk et al., 2013) in the last step. This means that even though the parameter estimates are correct, statistical inferences are not. When underestimated standard errors are used the confidence intervals will be too narrow and significance tests overly optimistic, thus increasing the probability of Type I error. At the same time, using overestimated standard errors leads to loss of power. Considering the broad applications of three-step latent class modeling, this is an undesirable situation.

The problem of additional variance caused by using estimates from a previous step has been dealt with in the context of non-linear models (Carroll et al., 2006), and three-step structural equation modeling (Skrondal & Kuha, 2012; Oberski & Satorra, 2013), and econometric theory for two-stages least squares is already well-developed (Murphy & Topel, 1985). In this paper we apply the general theory of Gong and Samaniego (1981) to latent class modeling, noting similarities and differences with these other approaches.

In this article we introduce two correction methods that are based on the general theory of Gong and Samaniego (1981) and can account for the bias in the standard errors. We evaluate different possible estimators of the standard errors using Monte Carlo simulations showing how the optimal variance estimator depends on the class assignment method. We also provide advice which estimators to use in different situations in order to obtain correct inferences. Based on this study, the methods discussed have been made available to applied researchers in the syntax version of the software Latent GOLD 5.00 (Vermunt & Magidson, 2013).

Whereas most of this paper focuses on correct inferences using Vermunt's approach that conditions on the first step ML estimates (an approach that can be used in most practical situations), in the discussion we introduce the possibility of Bayesian inference, showing how the uncertainty about the first step parameters can be accounted for in the last step using multiple imputation. The Bayesian approach can be useful for instance in situations where model uncertainty is high, or sample size is low and there are strong priors available.

The structure of the paper is as follows: in Section 3.2 we introduce the bias-adjusted three-step latent class analysis. Next section 3.3 presents possible variance estimators of this model. Section 3.4 then evaluates and compares the performance of these different

variance estimators in a simulation study. Section 3.5 revisits McCutcheon's (1985) analysis of how education and age groups differ in their tolerance. While the author initially used the uncorrected three-step approach, we show how inferences change using the corrections we propose. We conclude in Section 3.6, also showing directions for the Bayesian implementation of the methods we propose.

3.2 Bias-adjusted three-step latent class analysis

To model the relationship between a latent classification and external variables of interest without allowing the external variables to influence the classification, a three-step approach may be followed (Vermunt, 2010; Hagenaars, 1990):

Step one Using only the indicator variables, estimate a latent class model;

Step two Based on the first-step latent class model, create a new observed variable W that assigns to each unit its estimated latent class membership;

Step three Relate the estimated classification W to the external variables of interest.

Note that the assigned scores on W obtained in the second step do not correspond exactly to the true values unless the classification is perfect. Thus, classification error is introduced. As a consequence the parameter estimates of the third-step model will be biased, since a model with variables with measurement error is estimated (Bolck et al., 2004; Hagenaars, 1990). However, a specific of this model is that the amount of error introduced in step two is known. Thus, the step three model can be augmented to account for this known classification error (Bolck et al., 2004; Vermunt, 2010). In the following we introduce in detail each of the steps, explaining how the step three model is corrected for classifictaion error.

Special attention is given to the possible variance estimators. In each step a choice can be made between Hessian or robust variance estimator, nevertheless it is not clear which is better. More importantly in the third step the variance estimators should also account for the additional variance due to the correction method implemented. We propose two correction methods, and later in the simulation study we cross the choice of Hessian or robust estimator with the choices of correction methods to give practical advice on which variance estimator to use.

Similar models, in which the amount of error in the proxy is known or the true value is approximated via multiple proxies, are available in political science literature (Blackwell, Honaker, & King, 2012) and in the literature on measurement error correction via latent variable models (Alwin, 2007; Fornell & Larcker, 1981; Skrondal & Kuha, 2012; Oberski & Satorra, 2013).

3.2.1 Step one: estimating a latent class model

The first step is a standard latent class analysis of K categorical indicator variables (McCutcheon, 1987; Goodman, 1974; Hagenaars, 1990). By indicator variables we un-

derstand, the observed variables used to define the LC model. Given a sample of n units, the observations $\mathbf{Y}_i$ are modeled as arising from T unobserved (latent) classes X,

$$P(\mathbf{Y}_i) = \sum_{t=1}^{T} P(X_i = t) P(\mathbf{Y}_i | X_i = t). \tag{3.1}$$

The $T-1$ unique latent class sizes (mixture proportions) will be denoted $P(X = t) = \rho_t$ and are the first set of parameters of the first-step model to be estimated.

Note that in most applications the number of latent classes is not known a priori, but can be selected based on a set of modification indices (AIC, BIC). While selecting the right number of classes is outside the focus of this paper, we recommend for those interested in this problem to refer to Nylund, Asparouhov, and Muthén (2007); Van der Heijden, 't Hart, and Dessens (1997); Sclove (1987).

Further, the responses of each unit to the K categorical indicator variables are usually assumed to be locally independent given the unit's latent class membership. The conditional probability of the i-th response given the latent class can then be written as a product of conditional item responses,

$$P(\mathbf{Y}_i | X_i = t) = \prod_{k=1}^{K} P(Y_{ik} | X_i = t) = \prod_{k=1}^{K} \prod_{r=1}^{R_k} \pi_{ktr}^{I(Y_{ik}=r)}, \tag{3.2}$$

where the indicator variable $I(Y_{ik} = r) = 1$ if subject i has response r on item k, and 0 otherwise. The last step assumes that conditional item responses are equal for all units and defines the $(K-1)KT$ unique probabilities $\{\pi_{ktr}\}$ as the second set of first-step model parameters to be estimated.

The first-step log-likelihood of the sample data L_1 follows by combining equations 3.1 and 3.2 and assuming independence of observations:

$$L_1(\boldsymbol{\theta}_1) = \sum_{i=1}^{N} \log P(\mathbf{Y}_i) = \sum_{i=1}^{N} \log \left[\sum_{t=1}^{T} \rho_t \prod_{k=1}^{K} \prod_{r=1}^{R_k} \pi_{ktr}^{I(Y_{ik}=r)} \right]. \tag{3.3}$$

The first-step parameter vector to be estimated $\boldsymbol{\theta}_1 = [\boldsymbol{\rho}, \boldsymbol{\pi}]$ collects the latent class sizes $\boldsymbol{\rho}$ and conditional item response probabilities $\boldsymbol{\pi}$. Sample estimates $\hat{\boldsymbol{\theta}}_1$ of the first-step parameters can be obtained by maximum-likelihood (ML). Usually expectation-maximization, a quasi-Newton method, or a combination of both is used to maximize the first-step likelihood in Equation 3.3.

The maximum-likelihood estimates are sample estimates and will contain sampling variance. Assuming that the first-step model in Equation 3.3 is correct, standard theory suggests that the sampling variance equals the inverse of the Fisher information (negative of the Hessian matrix):

$$\boldsymbol{\Sigma}_1^H = (-\mathbf{H})^{-1}, \tag{3.4}$$

where the Hessian matrix $\mathbf{H}$ is defined as the second derivative of the first-step data log-likelihood with respect to the first-step parameters, $\mathbf{H} = \partial^2 L_1 / \partial \boldsymbol{\theta}_1 \, \partial \boldsymbol{\theta}_1'$.

The first-step model may not be correct-for instance because the local independence assumption may not hold. If this misspecification is small, it will likewise have a small

effect on the first-step estimates $\hat{\boldsymbol{\theta}}_1$. However, misspecification then still affects standard errors and sampling variance. The robust or "sandwich" variance should then be used,

$$\boldsymbol{\Sigma}_1^R = \boldsymbol{\Sigma}_1^H \, \mathbf{B} \, \boldsymbol{\Sigma}_1^H, \tag{3.5}$$

where the "meat" of the sandwich, $\mathbf{B}$, is the average outer product of the case wise gradients (White, 1982). Although the robust variance estimator corrects for model misspecification, it will also lead to a loss of efficiency (Kauermann & Carroll, 2001). It is therefore not clear in practice whether $\boldsymbol{\Sigma}_1^H$ or $\boldsymbol{\Sigma}_1^R$ should be preferred.

In situations where the misspecification is strong using robust standard errors does not suffice. As King and Roberts (2012) highlight in these situations instead of relying on robust standard errors it is recommended to check where the misspecification is located, and correct for that. Some useful tools to check for misspecification are: the BVR statistics that checks whether there is residual association between two indicators after controlling for latent class membership (Vermunt & Magidson, 2013, p.72-73), or the EPC statistics, that shows how much the model parameters would change if one parameter was freed (Oberski & Vermunt, 2013). These statistics are especially useful to test whether there is a direct effect between an indicator and external variable or if there is residual association left between the indicators after controlling for class membership. If such effects exist they should be modeled in the step one model.

For simplicity of exposition in the following we restrict ourselves to models where all model assumptions hold, that is the conditional independence assumption of the indicators holds, and there are no direct effects of external variables on the indicators. Furthermore we assume that the number of classes is known. [2]

3.2.2 Step two: assignment of units to classes

After estimating the latent class model in the first step, a new variable W is created, assigning each unit to an estimated class. Following Bayes' rule, each unit's posterior probability of belonging to class t is

$$P(X_i = t|\mathbf{Y}_i) = \frac{P(X_i = t)P(\mathbf{Y}_i|X = t)}{P(\mathbf{Y}_i)}. \tag{3.6}$$

Sample estimates of the posterior probabilities $P(X_i = t|\mathbf{Y}_i)$ can be obtained by replacing class sizes $P(X = t)$ with $\hat{\rho}_t$, conditional probability $P(\mathbf{Y}_i|X = t)$ with $\prod \hat{\pi}_{ktr}$, and generally substituting elements of $\boldsymbol{\theta}_1$ in Equation 3.6 with their first-step sample estimates $\hat{\boldsymbol{\theta}}_1$. These estimates can be used in different ways to create an estimated class membership variable W (Vermunt, 2010). We introduce the two most widely known and applied assignment rules: modal and proportional assignment.

The modal assignment rule to generate a posterior classification W is the most widely used rule (Collins & Lanza, 2010, p. 72). Each unit is simply assigned the class label with the largest (modal) estimated posterior probability from Equation 3.6. Using modal

[2] In situations where there is model uncertainty in step one a Bayesian approach might be used, as introduced in the discussion. Note that uncertainty about the number of classes cannot easily be handled with the Bayesian approach, only uncertainty about direct effects.

assignment the value of $P(W_i = t|\mathbf{Y}_i) = 1$ is assigned for $P(X_i = t|\mathbf{Y_i}) > P(X_i = t^{'}|\mathbf{Y_i})$ for all $t \neq t^{'}$. For all other classes this value is set to 0, leading to a hard partitioning.

Proportional assignment, in contrast, is a soft partitioning method (Dias & Vermunt, 2008). For each unit, T records are first created, one for each latent class. The T values of W_i are then set equal to the posterior probabilities $P(X_i = t|\mathbf{Y}_i)$. The data matrix is therefore expanded to include T instead of one records for each of the n units, where the within-unit values of the class assignment variable W will act as weights in the third step of the analysis.

Irrespective of the assignment method used, the true (X) and assigned (W) class membership scores will differ. Classification errors are inevitable, even if the entire population were observed. The amount of classification errors will depend on the posterior classification and the assignment method chosen. After assignment, the assignment variable W will require correction for classification errors in the third step; therefore, the amount of error in it must first be calculated (Bolck et al., 2004).

Summing over all observed data patterns the amount of classification errors can be expressed as the posterior class membership conditional on the true value (Vermunt, 2010; Bakk et al., 2013),

$$P(W = s|X = t) = \frac{\frac{1}{N}\sum_{i=1}^{N} P(X_i = t|\mathbf{Y}_i)P(W_i = s|\mathbf{Y}_i)}{P(X = t)}. \tag{3.7}$$

Note that while for any assignment method used the general form of equation 3.7 is the same, the values of $P(W_i = s|\mathbf{Y}_i)$ will differ per assignment method, and thus the amount of classification error $P(W = s|X = t)$ will also differ per assignment method. For example, $P(W_i = s|\mathbf{Y}_i)$ is either 0 or 1 using modal assignment, and with proportional assignment $P(W_i = s|\mathbf{Y}_i) = P(X_i = s|\mathbf{Y}_i)$. As we will show later this difference is not problematic, it just reflects that the amount of classification error depends on the assignment method used.

The classification error can be re-expressed on the logit scale as follows:

$$P(W = s|X = t) = \frac{\exp(\gamma_{st})}{\sum_{s=1}^{T} \exp(\gamma_{st})}, \tag{3.8}$$

where

$$\gamma_{st} = \log\left[\frac{P(W = s|X = t)}{P(W = t|X = t)}\right].$$

Note that the logistic γ_{st} parameters do not constitute free parameters but follow as a function of the first-step results and the assignment rule chosen.

We collect the γ_{st} parameters in the vector $\boldsymbol{\theta}_2$, with sample estimates $\hat{\boldsymbol{\theta}}_2$, calculated directly from $\hat{\boldsymbol{\theta}}_1$. These logistic effects of the true latent class on the estimated classification W are later needed to correct for classification error. Since these logit coefficients are calculated from the uncertain first-step estimates, they are themselves uncertain. Their

sampling variance $\mathbf{\Sigma}_2$ can be obtained using the delta method from the variance of the first step model (Oehlert, 1992):

$$\Sigma_2 = \left(\frac{\partial\theta_2}{\partial\theta_1}\right)\Sigma_1\left(\frac{\partial\theta_2}{\partial\theta_1}\right)' . \tag{3.9}$$

Either of the Σ_1 estimators discussed above can be plugged in to the formula, leading to an observed Hessian based (Σ_2^H) or robust (Σ_2^R) variance estimator of the second step parameters.

3.2.3 Step three: relating estimated class membership to covariates

In the third step the assigned classification W is related to a vector of covariates, $\mathbf{Z}$, say, while also correcting for classification error in W. Logistic regression of W on $\mathbf{Z}$ may appear to be an obvious solution, but would yield biased estimates due to classification errors in W. In effect, the relationship with the error-prone W is modeled, where the relationship with the true but unobserved latent class variable X is of interest, leading to measurement error effects on the parameter estimates (Bolck et al., 2004).

Bolck et al. (2004) showed how the $P(X = t|\mathbf{Z}_i)$, and $P(W = s|\mathbf{Z}_i)$ are related to each other, namely that the $P(W = s|\mathbf{Z}_i)$ can be written as a weighted sum of the latent classes given the covariates,with the classification error probabilities as the weights:

$$P(W = s|\mathbf{Z}_i) = \sum_{t=1}^{T} \underbrace{P(X = t|\mathbf{Z}_i)}_{\text{free}} \underbrace{P(W = s|X = t)}_{\text{fixed}} . \tag{3.10}$$

Details of the derivation are available in Bolck et al. (2004). Equation 3.10 can be seen as a latent class model with W as a single indicator that is fixed to the "known" classification error probabilities $P(W = s|X = t)$, as defined in Equation 3.8 (Vermunt, 2010). This means that relating the estimated membership to covariates while correcting for classification errors can be achieved by using standard latent class software that allows the user to fix classification error parameters to those obtained in the second step. This model is composed of two parts: (1) the structural part, i.e. the model of interest for $P(X = t|\mathbf{Z}_i)$, relating the latent class membership to the vector of external variables and (2) the measurement part $P(W = s|X = t)$ fixed to the parameter values estimated in step 2, as shown in Equation 3.7.

Denoting by Z_{iq} the value of subject i on one of the Q covariates, the structural part of the model can be parametrized by means of a multinomial logistic regression model,

$$P(X = t|\mathbf{Z}_i) = \frac{\exp(\beta_{0t} + \sum_{q=1}^{Q} \beta_{qt} Z_{iq})}{\sum_{s=1}^{T} \exp(\beta_{0t} + \sum_{q=1}^{Q} \beta_{qs} Z_{iq})} . \tag{3.11}$$

Although we only present the third-step model with predictors of latent class membership here, Bakk et al. (2013) showed how the correction method can be used for a wider

class of models, including models where the class membership is a predictor of a distal outcome variable, or with multiple latent variables. For the measurement part the logistic parametrization can be used as defined in Equation 3.8.

The parameters of interest are the logistic regression coefficients β_{qt}, gathered in the vector $\boldsymbol{\theta}_3$. Consistent estimates $\hat{\boldsymbol{\theta}}_3$ can be obtained by maximizing the third-step log-likelihood (Vermunt, 2010),

$$L_3(\boldsymbol{\theta}_3 \,|\, \boldsymbol{\theta}_2 = \hat{\boldsymbol{\theta}}_2) = \sum_{i=1}^{N} \sum_{s=1}^{T} P(W = s|\mathbf{Y}_i) \log \sum_{t=1}^{T} P(X = t|\mathbf{Z}_i) P(W = s|X = t). \tag{3.12}$$

Thus, in the third step, the logistic regression coefficients, contained in the third-step parameter vector $\boldsymbol{\theta}_3$, are freely estimated, while the classification errors of the class membership variable W as a measure of X, contained in the second-step parameter vector $\boldsymbol{\theta}_2$, are held fixed at their sample maximum-likelihood estimates, $\boldsymbol{\theta}_2 = \hat{\boldsymbol{\theta}}_2$. The third-step ML estimates can therefore be seen as conditional estimates $(\hat{\boldsymbol{\theta}}_3 | \, \boldsymbol{\theta}_2 = \hat{\boldsymbol{\theta}}_2)$.

3.3 Variance of the third-step estimates

Although the third-step maximum-likelihood estimates $\hat{\boldsymbol{\theta}}_3$ are consistent, their sampling variance now contains two sources of variation: that variation due to estimation at the third step, and that carried over from the first step. Ignoring the second source of variance will lead to an underestimation of the standard errors, as the results of previous simulation studies showed (Vermunt, 2010; Bakk et al., 2013).

In the following we introduce two correction methods to account for this additional uncertainty. We also highlight a special problem of proportional assignment that needs to be solved regardless of the choice made for correction for uncertainty in the variance estimator.

To see why underestimation occurs, write the variance of the third-step estimate as conditional on the second step (Oberski & Satorra, 2013):

$$\boldsymbol{\Sigma}_3^* \equiv \mathrm{Var}(\hat{\boldsymbol{\theta}}_3) = \mathrm{E}_{\theta_2}[\mathrm{Var}(\hat{\boldsymbol{\theta}}_3 | \, \boldsymbol{\theta}_2)] + \mathrm{Var}_{\theta_2}[\mathrm{E}(\hat{\boldsymbol{\theta}}_3 | \, \boldsymbol{\theta}_2)]. \tag{3.13}$$

The first term in Equation 3.13 corresponds approximately to the usual variance calculations obtained after fixing parameters in the third step,

$$\mathrm{E}_{\theta_2}[\mathrm{Var}(\hat{\boldsymbol{\theta}}_3 | \, \boldsymbol{\theta}_2)] \approx \boldsymbol{\Sigma}_3, \tag{3.14}$$

where $\boldsymbol{\Sigma}_3$ may, again, be estimated as the inverse third-step Fisher information or with the robust variance estimator. This is the basis for standard errors currently given by standard latent class analysis software when performing three-step analysis.

In the case of proportional assignment, each unit has several cases associated with it. Simulation studies by Vermunt (2010) and Bakk et al. (2013) found that using the third-step Hessian matrix to obtain an estimator of $\boldsymbol{\Sigma}_3$, standard errors were underestimated for modal assignment but overestimated for proportional assignment, a phenomenon that can be explained by the duplication of records present in proportional assignment. To

correct the standard errors for this duplication, $\boldsymbol{\Sigma}_3$ must be estimated with the well-known "complex sampling" (clustered) robust variance estimator (Wedel, Ter Hofstede, & Steenkamp, 1998), which will be denoted $\boldsymbol{\Sigma}_3^R$. Using this estimator we expect the standard error estimates to be down-weighted, because the sum of square of weights is always smaller then one with proportional assignment.

The second term in Equation 3.13 can be obtained by a first-order Taylor expansion (Gong & Samaniego, 1981; Oberski & Satorra, 2013),

$$\mathrm{Var}_{\theta_2}[\mathrm{E}(\hat{\boldsymbol{\theta}}_3|\,\boldsymbol{\theta}_2)] \approx \left(\frac{\partial\,\boldsymbol{\theta}_3}{\partial\,\boldsymbol{\theta}_2}\right)\boldsymbol{\Sigma}_2\left(\frac{\partial\,\boldsymbol{\theta}_3}{\partial\,\boldsymbol{\theta}_2}\right)', \tag{3.15}$$

where an estimate of $\boldsymbol{\Sigma}_2$ is available from the second step, and $\partial\,\boldsymbol{\theta}_3\,/\partial\,\boldsymbol{\theta}_2$ can be obtained using implicit function theorem:

$$\frac{\partial\,\boldsymbol{\theta}_3}{\partial\,\boldsymbol{\theta}_2} = \left(-\frac{\partial^2 L_3}{\partial\,\boldsymbol{\theta}_3\,\partial\,\boldsymbol{\theta}_3'}\right)^{-1}\frac{\partial^2 L_3}{\partial\,\boldsymbol{\theta}_3\,\partial\,\boldsymbol{\theta}_2'} \equiv -\,\mathbf{H}_3^{-1}\,\mathbf{C}, \tag{3.16}$$

which thus requires obtaining the second derivatives of the third-step log likelihood towards the free parameters ($\mathbf{H}$) and towards the free parameters with respect to the fixed parameters ($\mathbf{C}$). Therefore, the third-step variance defined in Equation 3.13 can be written as the sum of two positive-definite terms,

$$\boldsymbol{\Sigma}_3^* = \boldsymbol{\Sigma}_3 + \mathbf{H}_3^{-1}\,\mathbf{C}\,\boldsymbol{\Sigma}_2\,\mathbf{C}'\,\mathbf{H}_3^{-1}\,. \tag{3.17}$$

If a second-order Taylor expansion is used instead of Equation 3.15, an additional term results (Gong & Samaniego, 1981, Theorem 2.2), leading to

$$\boldsymbol{\Sigma}_3^{**} = \boldsymbol{\Sigma}_3 + \mathbf{H}_3^{-1}(\mathbf{C}\,\boldsymbol{\Sigma}_2\,\mathbf{C}' - \mathbf{C}\,\mathbf{H}_2^{-1}\,\mathbf{R}' - \mathbf{R}\,\mathbf{H}_2^{-1}\,\mathbf{C}')\,\mathbf{H}_3^{-1}, \tag{3.18}$$

where the $\mathbf{R}$ matrix is the outer product of the case-wise gradients of the first and third-step models, $\mathbf{R} = (\partial L_3/\partial\,\boldsymbol{\theta}_3)'(\partial L_1/\partial\,\boldsymbol{\theta}_1)$. However, perhaps surprisingly, this extra term vanishes as the sample size increases: provided the first-step estimates are consistent, asymptotically $\mathbf{R} = \mathbf{0}$ (Parke, 1986). Therefore, the two variance estimators are equal in large samples, $\boldsymbol{\Sigma}^{**} \stackrel{a}{=} \boldsymbol{\Sigma}^*$, although they may not be equal in small samples. In small samples it is possible that $\boldsymbol{\Sigma}^*$ will overestimate the standard errors of the third-step estimates, although this overestimation should decrease as sample size increases; on the other hand, the calculation of the extra terms in $\boldsymbol{\Sigma}^{**}$ may add considerable effort and instability to the standard errors.

Whether $\boldsymbol{\Sigma}^*$ or $\boldsymbol{\Sigma}^{**}$ is the more appropriate variance estimate is therefore unclear. Furthermore, it can be concluded from the preceding discussion that at each step a range of possible choices of variance estimators exist. The following section investigates how combinations of these different choices perform and which, if any, of the standard error corrections is likely to be necessary in practice.

Table 3.1: Possible variance estimators of the third step model

Final	Components	
	2^{nd} step	3^{rd} step
Uncorrected ($\mathbf{\Sigma}_3$)	-	Hessian ($\mathbf{\Sigma}_3^H$)
	-	Robust ($\mathbf{\Sigma}_3^R$)
1^{st} order correction ($\mathbf{\Sigma}_3^*$)	Hessian ($\mathbf{\Sigma}_2^H$)	Hessian ($\mathbf{\Sigma}_3^H$)
	Hessian($\mathbf{\Sigma}_2^H$)	Robust ($\mathbf{\Sigma}_3^R$)
	Robust($\mathbf{\Sigma}_2^R$)	Robust ($\mathbf{\Sigma}_3^R$)
2^{nd} order correction ($\mathbf{\Sigma}_3^{**}$)	Hessian($\mathbf{\Sigma}_2^H$)	Hessian ($\mathbf{\Sigma}_3^H$)
	Hessian ($\mathbf{\Sigma}_2^H$)	Robust ($\mathbf{\Sigma}_3^R$)
	Robust ($\mathbf{\Sigma}_2^R$)	Robust ($\mathbf{\Sigma}_3^R$)

3.4 Monte Carlo simulation

3.4.1 Design

In order to see which variance estimator performs the best, we crossed the choice of variance estimators (for $\mathbf{\Sigma}_2$ and $\mathbf{\Sigma}_3$: observed Hessian based or robust) with the options for correcting for uncertainty ($\mathbf{\Sigma}_3$ - uncorrected, $\mathbf{\Sigma}_3^*$ first order and $\mathbf{\Sigma}_3^{**}$ second order correction) for both modal and proportional assignment. In the following table we summarize the different choices of the variance estimators compared.[3]

As used in Table 3.1 the 1^{st} order correction, $\mathbf{\Sigma}_3^*$ is defined in Equation 3.17 and the 2^{nd} order correction, $\mathbf{\Sigma}_3^{**}$ in 3.18, and $\mathbf{\Sigma}_3$ is the variance of the free parameters ignoring the additional uncertainty attributable to the fixed parameter values. In reporting the simulation study results and real data example we use the term $\mathbf{\Sigma}_3^R$ in case of proportional assignment for the complex sampling variance estimator (Wedel et al., 1998), and for modal assignment for the sandwich estimator as defined by White (1982). All in all we investigate 8 variance estimators for each of the two assignment methods separately.

The need for the uncertainty correction is expected to depend on the amount of uncertainty about the model parameters, that we varied by changing sample size and separation between classes.

As population model we chose a LCA model with three classes measured by six dichotomous indicators, and regressed on three numerical covariates (each with five categories: 1-5). The first class is likely to give positive response on all 6 items, class two has a high probability of a positive response on the first three items, and negative response on the other three items. In class three all items have a high probability of a negative answer. We manipulated the separation between classes by changing the size of the conditional probability of the indicators given the classes. The two levels of separation we used for the probability of a positive answer are .80 and .90, corresponding to entropy R^2 values of .65 and .90. We chose the following sample sizes: 500, 1000, 2000. Thus in total we had six conditions of combinations of sample size and separation between classes, in

[3]The simulation set up is available in the dataverse replication material of this article: Study Global Id: doi:10.7910/DVN/24497 (v1)

Table 3.2: Parameter estimates and their standard deviation (sd) for all parameters averaged over all conditions for all estimators

		Modal		Proportional		One-Step	
Value	True	Estimate	sd	Estimate	sd	Estimate	sd
β_{12}	-2.00	-1.98	0.30	-1.97	0.28	-2.07	0.30
β_{13}	1.00	1.00	0.12	1.00	0.11	1.01	0.11
β_{22}	1.00	1.00	0.17	0.98	0.16	1.02	0.18
β_{23}	0.00	0.00	0.08	0.00	0.07	0.00	0.08
β_{32}	0.00	0.00	0.11	0.00	0.11	0.00	0.11
β_{33}	0.00	0.00	0.07	0.00	0.07	0.00	0.07

which the performance of all eight variance estimators was compared for both modal and proportional assignment. For each condition 500 replications were used.

Using the first class as reference category we set the logit parameters of covariate effects on latent classes to -2 (β_{12}) and 1 (β_{13}) for the effect of Z_1 on X. Where we use the first subscript for Z, and the second subscript for X, as such for example β_{13} stands for the effect of Z_1 on the third class. Te effect of Z_2 on X is set to 1 (β_{22}) and 0 (β_{23}), and to 0 for both parameters (β_{32}, β_{33}) for the effect of Z_3 on X. The intercepts were set to values yielding equal class sizes.

Two measures were used to compare the performance of the variance estimators. We compared the coverage rate over replications to a nominal 95 % rate, and the average standard errors (se) across replications to the standard deviation (sd) across replications. For a well performing standard error estimator we expect the se/sd to be 1. Also the coverage rate should be 95 %, which is the nominal coverage rate used.

We used the computer programs Latent GOLD (Vermunt & Magidson, 2013) and R (Venables, Smith, the R Core Team, 2013) to run the analysis.

3.4.2 Simulation results

First we compare the parameter estimates and standard deviation across replications obtained with the three-step approach with the two assignment methods and the one-step approach in order to see whether the three-step estimates are comparable with regard to parameter bias and efficiency to the estimates obtained using the one-step approach. In Table 3.2 we report the mean parameter estimates over all replications, and the standard deviation across replications for all three estimation methods. On average the parameter bias is low with all three estimators for all the parameters. We compared the efficiency of the parameter estimators by comparing the standard deviation across replications. As we can see in Table 3.2 the standard deviations of all parameters are very close to each other with the three methods. These results are in accordance with previous simulation studies (Bakk et al., 2013; Vermunt, 2010), and show that the three-step approach can be used without loss of efficiency or parameter bias.

Given the unbiased parameter estimates reported in Table 3.2 in the following we restrict the discussion only to the variance estimators of the third-step model. Let us first

Table 3.3: Comparison of the different variance estimators averaged across all conditions for one parameter, β_{13} for modal and proportional assignment separately

Final	Components		Modal			Proportional		
	2^{nd} step	3^{rd} step	se	se/sd	coverage	se	se/sd	coverage
$\mathbf{\Sigma}_3$	-	$\mathbf{\Sigma}_3^H$	.11	0.95	.95	.12	1.08	.97
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^H$	.12	1.03	.96	.13	1.14	.98
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^H$	.12	1.04	.96	.13	1.12	.97
$\mathbf{\Sigma}_3$	-	$\mathbf{\Sigma}_3^R$	.12	0.97	.95	.11	0.99	.95
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^R$	.12	1.04	.96	.12	1.05	.96
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^R$	.13	1.05	.96	.12	1.03	.96
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^R$	$\mathbf{\Sigma}_3^R$	.13	1.05	.96	.12	1.06	.96
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^R$	$\mathbf{\Sigma}_3^R$	.13	1.06	.96	.12	1.04	.96

Note: $\mathbf{\Sigma}_3^*$ is the 1^{st} and $\mathbf{\Sigma}_3^{**}$ the 2^{nd} order correction, as defined in equation 17 and 18, and $\mathbf{\Sigma}^H$ and $\mathbf{\Sigma}^R$ are the Hessian based and robust estimators

look on the results averaged across all conditions of sample size and separation between classes, that are reported in Table 3.3 for one parameter ($\beta_{13} = 1.00$). The results for the other parameters are very similar.

For modal assignment, as can be seen in Table 3.3 the two uncorrected standard error estimators that do not account for the additional uncertainty (Σ_3^H and Σ_3^R) underestimate the variance (the se/sd is .95 for Σ_3^H, and .97 for Σ_3^R). Using either of the correction methods ($\mathbf{\Sigma}_3^*$ or $\mathbf{\Sigma}_3^{**}$) improves the results for both Hessian based and robust estimator. Comparing the first and second order corrections ($\mathbf{\Sigma}_3^*$, $\mathbf{\Sigma}_3^{**}$) to each other we see that the standard error estimates obtained with the later are slightly higher irrespective of the choice for robust or Hessian based estimator. When comparing observed Hessian based to robust estimator for the modal assignment, we see that the standard errors obtained with the later are somewhat larger, thus less efficient. The differences are small, as can be seen from the coverage rate, which is almost the same with all estimators.

Next, looking on the results for proportional assignment, we see, that as hypothesized the standard error estimates obtained with the observed Hessian based estimator overestimate the standard error for all three estimators ($\mathbf{\Sigma}_3, \mathbf{\Sigma}_3^*, \mathbf{\Sigma}_3^{**}$). This can be seen from both the se/sd (which is higher than 1 for all three estimators) and coverage rate (that is .97 for $\mathbf{\Sigma}_3$ and $\mathbf{\Sigma}_3^{**}$ and .98 for $\mathbf{\Sigma}_3^*$). Using the robust variance estimator in the third step improves the results (the se/sd using $\mathbf{\Sigma}_3^R$ is .99, and using the correction methods this value becomes slightly larger then 1). Similarly to modal assignment we can see, that using robust variance estimator in the first step yields larger standard error estimates.

Following we look separately into the results averaged over the different levels of separation between classes and the different sample size conditions. First the results averaged over the three sample sizes separately for the 2 separation levels are presented. For the condition with high separation between the classes (entropy R^2=.90) all variance estimators perform well. In case of modal assignment for all the variance estimators the ratio of the standard error to the standard deviation (se/sd) is between 1.00-1.02, and the coverage rate is 96 %, results that show that all standard error estimators perform well in

Table 3.4: Comparison of the different variance estimators across the three sample sizes,for the low separation levels for one parameter, β_{13} for modal and proportional assignment separately

Final	Components		Modal			Proportional		
	2^{nd} step	3^{rd} step	se	se/sd	coverage	se	se/sd	coverage
$\mathbf{\Sigma}_3$	-	$\mathbf{\Sigma}_3^H$	.12	0.91	.93	.14	1.11	.98
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^H$	.14	1.03	.96	.15	1.21	.98
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^H$	.14	1.05	.96	.15	1.18	.98
$\mathbf{\Sigma}_3$	-	$\mathbf{\Sigma}_3^R$	.13	0.93	.94	.12	0.97	.95
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^R$	.14	1.05	.96	.14	1.08	.97
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^R$	.14	1.06	.96	.13	1.05	.96
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^R$	$\mathbf{\Sigma}_3^R$	.15	1.07	.96	.14	1.10	.97
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^R$	$\mathbf{\Sigma}_3^R$	.15	1.08	.96	.14	1.07	.96

Note: $\mathbf{\Sigma}_3^*$ is the 1^{st} and $\mathbf{\Sigma}_3^{**}$ the 2^{nd} order correction, as defined in equation 17 and 18, and $\mathbf{\Sigma}^H$ and $\mathbf{\Sigma}^R$ are the Hessian based and robust estimators

this condition. The same holds for all standard error estimates for proportional assignment that are based on the robust variance estimator. As such we do not present these results in more detail, but move toward the discussion of the low separation condition, where we see more variability.

In Table 3.4 the results averaged over the three sample sizes for the low separation condition are presented. For modal assignment the uncertainty uncorrected standard error estimate ($\mathbf{\Sigma}_3$) underestimates the standard error with both Hessian based and robust estimators (se/sd is .91 and .93 while coverage rate is .93 and .94 for Σ_3^H and Σ_3^R respectively). Using either of the correction methods ($\mathbf{\Sigma}_3^*, \mathbf{\Sigma}_3^{**}$) the se/sd becomes slightly larger then 1, and the coverage rate increases to 96%, for both the Hessian based and robust estimators. When comparing the observed Hessian based estimates to the robust estimates we can see that the later obtains slightly larger standard error estimates.

For proportional assignment we can see that the variance estimators that use the observed Hessian in the third-step overestimate the standard error. Using the robust variance estimator in the third step decreases the variance. Once this is used, the difference between the standard error estimates is small ($\mathbf{\Sigma}_3$ obtains se/sd 0.97,with $\mathbf{\Sigma}_3^*$ this is 1.08, and with $\mathbf{\Sigma}_3^{**}$ 1.05). Using robust variance estimator in the first step increases the standard error estimates.

Next in Table 3.5 we present the standard error estimates averaged over the two separation levels separately for the three sample size conditions. For modal assignment we can see that in the small sample size condition the uncorrected standard error estimates are underestimated (se/sd 0.87 for Σ_3^H and 0.90 for Σ_3^R and coverage rate 0.92 and 0.93 respectively), but using any of the correction methods this values get closer to 1. Comparing the first and second order correction ($\mathbf{\Sigma}_3^*$, $\mathbf{\Sigma}_3^{**}$) we see that the standard error estimates obtained with the later are slightly larger irrespective of the choice of $\mathbf{\Sigma}_3$. The same tendencies can be seen in the larger sample size conditions as well, though in the 2000 sample size condition it can be seen, that on average the uncorrected standard error

estimates are the same as the corrected ones with the precision of 2 decimals. Comparing the Hessian based estimators to the robust ones we see that the later ones are somewhat larger. Using proportional assignment the same tendencies can be observed, once in the third step the robust standard error is used.

In summary it can be said, that in conditions where the uncertainty about the fixed parameters is high (that is low separation between classes and/or low sample sizes) the use of the uncertainty correction is needed.[4] It can be seen that the difference with the results using $\boldsymbol{\Sigma}_3^*$ and $\boldsymbol{\Sigma}_3^{**}$ is low, thus the use of the first order correction is recommended, because it needs less calculations. With regard to the choice of Hessian or robust variance estimator we see that in case of proportional assignment this choice is important. With proportional assignment the use of robust estimator is recommended for all situations, while for modal assignment this choice is not so relevant.

3.5 Example application

We now show how correcting for parameter bias in the three-step latent class analysis makes a difference for substantive conclusions. For this purpose we re-analyze the often-cited example of latent class analysis in political science, McCutcheon (1985)'s assessment of how age and education groups differ in their tolerance towards out-groups. In addition, we illustrate how the different choices of standard error discussed above can affect results.

The question of who is more intolerant originated with Stouffer (1955). His analysis, conducted at the height of the McCarthy era, focused on citizens' tolerance for communists: should they be allowed basic democratic rights to free speech according to the public? McCutcheon (1985) re-assessed this question by including other groups in the questionnaire besides communists, namely atheists, homosexuals, militarists, and racists. He then used a latent class model to integrate respondents' tolerance for these different groups into a single categorical latent variable that represents each person's "tolerance for nonconformity". After estimating this model (step 1) and assigning a "tolerance" classification to each respondent (step 2), McCutcheon then regressed the categorical "tolerance" assignment on age cohort and educational attainment (step 3). This commonly cited example of a latent class analysis in political science is therefore in fact a bias-uncorrected three-step analysis.

Here we examine how McCutcheon (1985)'s conclusions might change when the necessary bias corrections (Vermunt, 2010) are applied to the third step, predicting "tolerance" from age cohort and education. We also show how the standard errors of this regression differ over the various choices of standard error estimator described in the preceding sections.

The original data are obtained from the 1976 and 1977 General Social Survey (GSS), which are publicly available[5]. Each of the 2689 respondents (nr. of respondents obtained

[4] This tendency can be seen in more detail in the Appendix in Tables B1 and B2, which show that once the uncertainty decreases (either by a stronger effect of classes on indicators or higher sample size) the effect of the corrections is lower, but when the uncertainty is high the corrections make a big difference.

[5] `https://www.icpsr.umich.edu/icpsrweb/ICPSR/series /28/studies/7398?archive=ICPSRsort`, also available in the dataverse replication material of this article: Study Global Id: doi:10.7910/DVN/24497 (v1)

after listwise deletion was applied) answered the following questions on communists, atheists, homosexuals, militarists, or racists:

- "Suppose this _____ wanted to make a speech in your community. Should he be allowed to speak?" (**Yes**/No)
- "Should such a person be allowed to teach in a college or university, or not?" (**Yes**/No)
- "Suppose he wrote a book which is in your public library. Somebody in your community suggests that the the book should be removed from the library. Would you favor removing it or not?" (Yes/**No**)

The bold-faced answers are those indicating tolerance. McCutcheon coded a respondent as "tolerant" towards a group if all three of these bold-faced answers were given, and "intolerant" otherwise. This yields five binary indicators of general tolerance (one for each group).

The first step is to fit local independence models with a successively increasing number of latent classes. Resulting model fit statistics are shown in Table 3.6. This Table shows that the AIC selects the four-class model, while the BIC selects the three-class model. Looking at the absolute model fit of the four-class model, however, it is clear that both the likelihood ratio as well as the largest bivariate residual (BVR; see Vermunt and Magidson (2013)) indicate residual dependencies between the indicators in the three-class model. We therefore follow McCutcheon (1985) and select the four-class model, which fits the data well. None of the residual dependencies between the observed variables are substantively large or statistically significant in the four-class model, thus we can reasonable assume that there is no residual variance left between the indicators.

Estimates of class sizes and conditional probabilities obtained from the four-class model are shown in Table 3.7. The first row of this table indicates the labels given to the four classes. These labels are based on the pattern of estimated conditional probabilities given on the subsequent rows. For example, the first class, to which 56% ($\pm 2\%$) of respondents are estimated to belong, exhibits low probabilities of tolerance for all groups, ranging between 0.03 for atheists to 0.13 for homosexuals (both ± 0.01). Therefore this latent class was labeled "intolerant". There are also classes of those respondents who are tolerant towards some groups and not others. Since these preferences appear to correspond to political ideology, McCutcheon labeled the classes "intolerant of right" and "intolerant of left" respectively. It should be noted, however, that this ideological intolerance is not symmetric: those who are intolerant of "right-wing" groups such as racists and militarists are more extreme in their opinion than those who are intolerant of "leftist" groups such as atheists or communists. The difference between these classes in their tolerance for homosexuals is not statistically significant ($z = -1.52$, $p = 0.064$).

While the latent class analysis of these five indicators (step 1) is interesting in itself, to Stouffer (1955) and McCutcheon (1985) the main substantive question was how age cohorts and educational groups differ in their overall tolerance. For this reason, each respondent was assigned to one of the four estimated classes using proportional assignment based on the latent class model (step 2). This assigned class variable is highly

convenient for further analysis: the analyst performing the latent class analysis may be separate from the researcher investigating substantive questions using the result. Thus, the researcher interested in the effect of cohort and education on tolerance need not have the same expertise as the latent class analyst. Furthermore, the definition of the latent class assignment has not been affected by the cohort and education variables, preventing circularity in the final results.

Following McCutcheon (1985), educational attainment was coded into three categories: those with fewer than twelve grades (1), those who completed high school (2), and those with more than twelve years of formal education (3). Birth cohort was coded into four categories: those born in or before 1914 (4), those born between 1915 and 1933 (3), those born between 1934 and 1951 (2), and those born after 1951 (1). We ran a multinomial regression of assigned tolerance on these covariates, corrected for misclassification error in the assignment and using the first categories of each variable as a reference category. The interaction effect of education $\times$ cohort turned out to be small and not statistically significant (Wald $= 12.2$ on 18 df, $p = 0.84$), and we therefore decided to exclude the interaction from further analysis. The resulting main effect estimates are shown as points with 95% confidence intervals in Figure 3.1.

Figure 3.1 summarizes the 15 multinomial logit coefficients and their standard errors from the uncorrected analysis (black dots) performed by McCutcheon (1985) and the corrected analysis (gray points). The effect size estimates show that the more educated and younger a person is, the more likely they are "tolerant". Looking at Figure 3.1 from bottom to top reveals a monotonic increase of logits with these two covariates. This applies to a lesser degree to the "intolerant to right" group, and to the "intolerant to left" group to an even lesser degree. This ordering in effect size from the rightmost to leftmost panel in Figure 3.1 is probably due to ideological differences between age and education groups: the younger and more educated were more likely to prefer the political left.

For comparisons between corrected and uncorrected analyses, the 15 coefficients from the corrected three-step analysis are shown below each estimate (gray points). It can be seen that there is a substantial difference between the corrected and the uncorrected estimates. To bring this point into perspective, Figure 3.2 shows the ratios of corrected to uncorrected point estimates. The most extreme case is the logit coefficient of the oldest cohort on being in the "intolerant to right" class, which increases almost twofold. Even the smallest correction entails a 20% larger coefficient, however. This emphasizes the practical significance of correcting for classification error in the class assignment: substantial bias will otherwise occur.

While the first-step sample size of 2689 is relatively large, the entropy R^2 of the "tolerance" latent classification is 0.71, falling short of the 0.90 our simulations identified as an indicator of "small" uncertainty about the classification error. The standard errors of this analysis may therefore benefit from correction for this uncertainty.

Does the correction to standard errors introduced in this paper make a difference for the results? Figure 3.1 shows that it does. First of all, qualitative differences can be observed for the effect of being in the younger cohort (2) on being "intolerant to left": this cohort is no longer deemed to differ significantly from the youngest cohort (1) when the correction is applied.

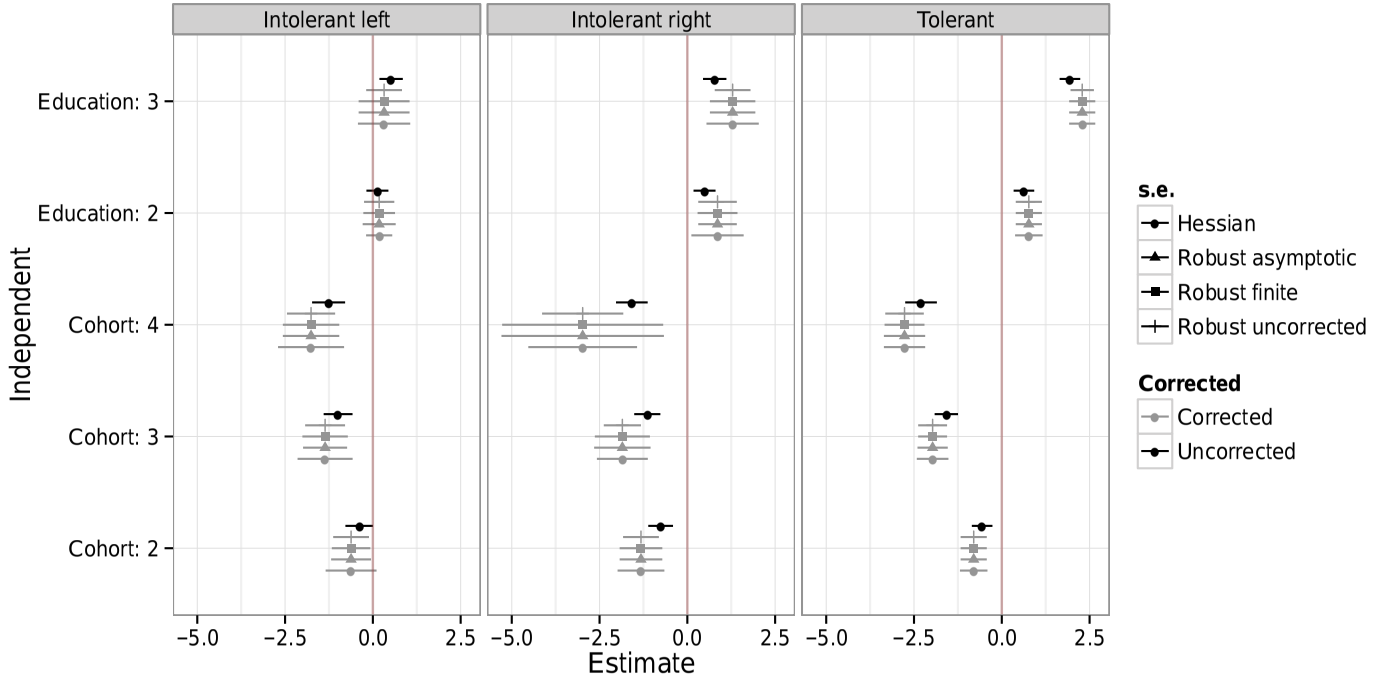

Figure 3.1: Multinomial effect size estimates of covariates on the tolerance classes from the third-step analysis. Shown are the dummy-coded effects of two education categories and three age cohorts on the four latent classes, using the first category as a reference. The uncorrected point estimate and 95% confidence interval is shown, below which the corrected point estimates with confidence intervals from four types of standard errors is given.

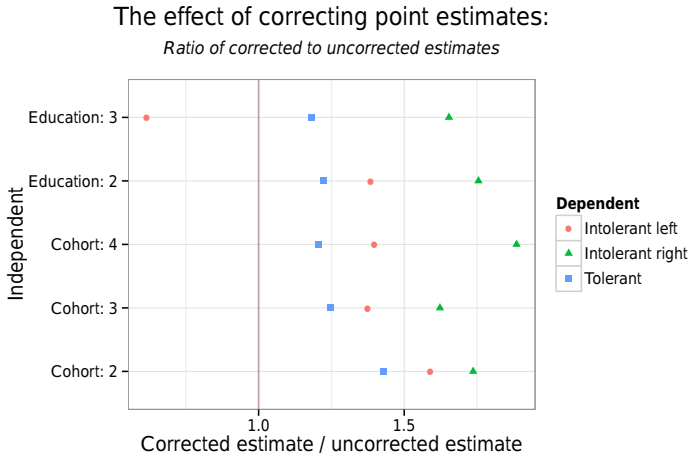

Figure 3.2: The relative size of the corrected point estimates compared with the uncorrected. Unity means the two estimates are the same, while 2 means the corrected point estimate is twice as large as the uncorrected estimate. The effects for different classes are indicated with different point shapes, given in the legend.

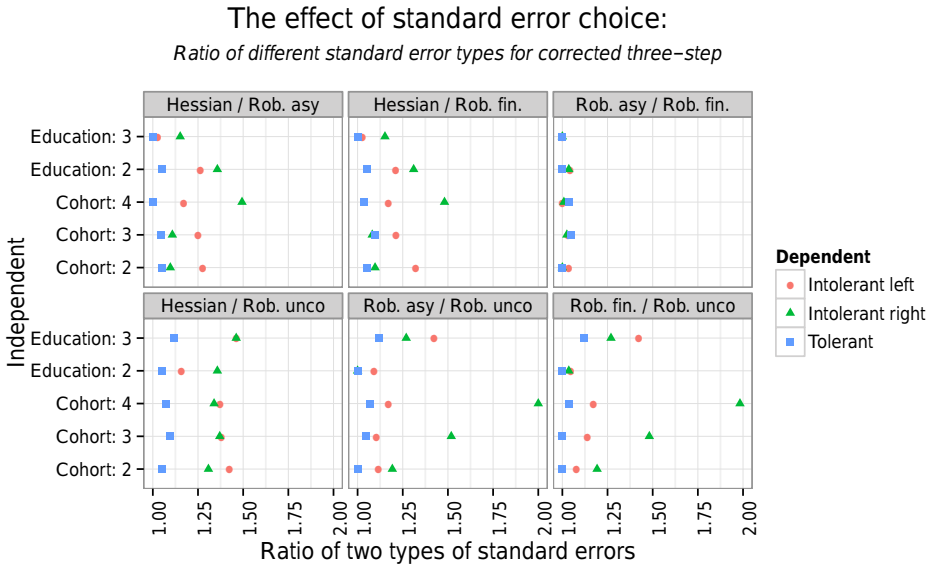

Figure 3.3: Comparison between the four different types of standard errors for the corrected procedure. Shown are the ratios of the larger standard error (se) to the smaller se for the six combinations shown in the panel titles.

Second, Figure 3.1 exhibits large quantitative differences in standard errors, particularly for the effects of being in the oldest cohort (4). Figure 3.3 shows the sizes of these standard errors relative to each other. For instance, the finite-sample robust standard error for the oldest cohort's effect on being "intolerant to right" is twice as large as the standard error without the correction discussed above (triangle in the bottom-right graph). Since the relative standard error must be squared to obtain a "misspecification effect" (Skinner, Holt, & Smith, 1989), this means that if the correction introduced here is ignored, the researcher will claim that the sample is four times more informative than it is. This clearly demonstrates the relevance of the corrections to standard errors introduced above.

3.6 Discussion and conclusion

Social scientists often aim to study the relationship of an unobserved classification with external variables. The "three-step" approach is common: a latent class model is fit (step 1), units are assigned to estimated classes (step 2), and the relationships of interest are studied using the assigned classes, for instance by multinomial regression (step 3). Three-step analysis potentially has several advantages over the "one-step" full-information maximum likelihood approach (Vermunt, 2010). Despite being common and attractive, this approach is also inconsistent — a problem solved by the bias-corrected three-step approach introduced by Bolck et al. (2004) and Vermunt (2010) in this journal and further generalized by Bakk et al. (2013).

Correct inferences about the relationships of interest, however, require not only consistent point estimates but also correct standard errors. In this article we show both analytically and by simulation that correct standard errors from the third step must incorporate the uncertainty about the classification error. We therefore provide in this paper the correct standard errors allowing for appropriate inferences, based on classic likelihood theory of Gong and Samaniego (1981). As a result of our study, the different standard error estimators discussed have been implemented in the standard latent class analysis software Latent GOLD 5.00 (Vermunt & Magidson, 2013), making the methods developed here directly available to applied researchers.

Moreover, we evaluate eight possible types of standard errors. Although these standard errors are asymptotically equivalent under model correctness, they may yield different results in finite samples. A Monte Carlo simulation study compared them and found that the correction to standard errors introduced here can make a large difference when uncertainty about the first-step parameters is substantial. On the other hand, when the uncertainty about fixed estimates is low, the standard error corrections are not needed. Low uncertainty about the classification error will occur with large first-step sample sizes and high entropy R^2 (high class separation). No substantial differences between inferences based on corrected versus uncorrected standard errors were found with first-step sample sizes above 2000 combined with entropy $R^2 > 0.90$. We also noted little difference between an asymptotic and finite-sample version of the corrected estimator. The asymptotic corrected standard error estimator (Oberski & Satorra, 2013), which is considerably easier to compute, is therefore recommended. Finally, we reproduced the finding of Vermunt (2010) that proportional assignment requires robust standard errors to account for the

replication of cases.

Considering these findings, bias correction and the choice of standard errors can make a difference for substantive conclusions. Reanalysis of an example bias-uncorrected three-step analysis from the political science literature (McCutcheon, 1985) clearly demonstrated this effect. The logistic regression coefficients that give the strength of the relationship between being in the "intolerant to the right" class and all age and education categories are between one-and-a-half and twice as large as their uncorrected counterparts, for instance. The correction for classification error need not always increase estimated relationships: one of the coefficients of interest is much lower after correction. "Qualitative" differences also occur after correcting both the point estimates and the standard errors: the younger cohort's logistic coefficient for "intolerant to left" is not statistically significant in the uncorrected analysis, but is so after correction. Conversely, the highly educated group's coefficient is statistically significant in the uncorrected analysis, but not after correction. This demonstrates the importance of the corrections, both for point estimates and standard errors.

A limitation of our study is that we restricted ourselves to situations where the separation between classes is relatively good (entropy $R^2 =$ 60 or higher). We did so because previous research showed that in situations where the entropy is lower than this, the step three methods obtain biased estimates. This is due to the fact that in step one the classification error is underestimated, and thus over-optimistic correction terms are used (Vermunt, 2010, Bakk et al., 2013).

A further limitation of our study is that we assumed an (approximately) correct model can be found. That is, we assume that mostly sampling variance drives the first-step model uncertainty. For this reason, model checking in the first step is essential. A possible alternative approach would be to obtain point and uncertainty estimates under model uncertainty, after which these may be propagated to the third step as described above.

A completely different approach is the Bayesian multiple imputation framework (Rubin, 1987). In this framework, the first step is to formulate a Bayesian latent class model, the second to obtain M multiple draws from the latent class distribution, and the third to estimate M regression models, averaging the M parameter estimates and using the rules described by (Schafer, 1997) to correctly obtain standard errors. The idea of using Bayesian data augmentation to estimate the conditional distribution of a latent variable X given predictors Z was introduced by Tanner and Wong (1987). Applications of this idea can be found in the "plausible values" literature for continuous latent variables (Mislevy, 1988), as well as the method of "pseudo-class draws" (Bandeen-Roche et al., 1997; Wang, Brown, & Bandeen-Roche, 2005; Asparouhov & Muthén, 2014).

However, the same inconsistency problems that plague the uncorrected three-step method also affect the Bayesian multiple imputation approach. The key difference between, three-step analysis, plausible values, and pseudo-class draws on the one hand,and the Bayesian augmentation literature, on the other is the inclusion of the covariates in the first-step model. As shown by Tanner and Wong (1987), the multiple imputations of the latent variable X must be generated from $p(X|Y, Z)$, not just $p(X|Y)$ as is done in the three-step, plausible values, and pseudo-class draws procedures. Leaving out the predictor variables Z from the first-step imputations will therefore cause the same inconsistency as

is present in likelihood-based uncorrected three-step analysis. This was also shown by the simulations of Asparouhov and Muthén (2014, Table 1).

However, including Z in the first-step analysis partially defeats the purpose of the three-step procedure. Moreover, researchers performing complicated latent variable models to publish imputations for the broader research community (König, Marbach, & Osnabrügge, 2013) cannot possibly foresee all predictor variables Z that might someday be of interest.

The problem of inconsistency in the Bayesian multiple imputation approach caused by ignoring Z in the imputation model can in principle be solved by performing the bias correction described above in each imputation. The combined corrected point estimates will then be consistent for $p(X|Z)$. Afterwards, combining the resulting multiple corrected estimates will yield correct standard errors (Schafer, 1997). This method could substantially improve the quality of inferences from recent efforts in political science to publish multiple imputations of latent variables for further analysis such as Democracy (Treier & Jackman, 2008) or party positions (König et al., 2013). It should be noted that the proposed correction method is not limited to Bayesian multiple imputation, but can be used in any situation in which an integration or missing data problem is solved by simulation and is based on a (step-one) model that excludes some of the relevant variables. The resulting corrected "pseudo-class draws" or "plausible values" analysis is an interesting and potentially useful application of the presented three-step approach that warrants future study.

Table 3.5: Comparison of the different variance estimators averaged across the two separation levels,for the 3 sample sizes for one parameter, β_{13} for modal and proportional assignment separately

Final	Components		N=500			N=1000			N=2000		
	- 2^{nd} step	3^{rd} step	se	se/sd	coverage	se	se/sd	coverage	se	se/sd	coverage
					Modal						
$\boldsymbol{\Sigma}_3$		$\boldsymbol{\Sigma}_3^H$	.16	0.87	.92	.11	1.01	.95	.08	1.05	.96
$\boldsymbol{\Sigma}_3^*$	$\boldsymbol{\Sigma}_2^H$	$\boldsymbol{\Sigma}_3^H$	.17	0.96	.95	.12	1.08	.97	.08	1.13	.97
$\boldsymbol{\Sigma}_3^{**}$	$\boldsymbol{\Sigma}_2^H$	$\boldsymbol{\Sigma}_3^H$	.17	0.97	.95	.12	1.09	.97	.08	1.12	.97
$\boldsymbol{\Sigma}_3$		$\boldsymbol{\Sigma}_3^R$	.16	0.90	.93	.11	1.02	.95	.08	1.06	.96
$\boldsymbol{\Sigma}_3^*$	$\boldsymbol{\Sigma}_2^H$	$\boldsymbol{\Sigma}_3^R$	.18	0.98	.95	.12	1.08	.97	.08	1.11	.97
$\boldsymbol{\Sigma}_3^{**}$	$\boldsymbol{\Sigma}_2^H$	$\boldsymbol{\Sigma}_3^R$	.18	0.99	.95	.12	1.08	.97	.08	1.13	.97
$\boldsymbol{\Sigma}_3^*$	$\boldsymbol{\Sigma}_2^R$	$\boldsymbol{\Sigma}_3^R$	.18	1.01	.95	.12	1.10	.97	.08	1.11	.97
$\boldsymbol{\Sigma}_3^{**}$	$\boldsymbol{\Sigma}_2^R$	$\boldsymbol{\Sigma}_3^R$	.18	1.01	.95	.12	1.09	.97	.08	1.13	.97
					Proportional						
$\boldsymbol{\Sigma}_3$		$\boldsymbol{\Sigma}_3^H$	.17	0.99	.96	.12	1.15	.98	.08	1.20	.98
$\boldsymbol{\Sigma}_3^*$	$\boldsymbol{\Sigma}_2^H$	$\boldsymbol{\Sigma}_3^H$	.18	1.06	.96	.12	1.20	.98	.09	1.25	.98
$\boldsymbol{\Sigma}_3^{**}$	$\boldsymbol{\Sigma}_2^H$	$\boldsymbol{\Sigma}_3^H$	.18	1.04	.96	.12	1.18	.98	.09	1.23	.98
$\boldsymbol{\Sigma}_3$		$\boldsymbol{\Sigma}_3^R$	.16	0.92	.94	.11	1.04	.96	.08	1.08	.96
$\boldsymbol{\Sigma}_3^*$	$\boldsymbol{\Sigma}_2^H$	$\boldsymbol{\Sigma}_3^R$	.17	0.99	.95	.11	1.10	.97	.08	1.14	.97
$\boldsymbol{\Sigma}_3^{**}$	$\boldsymbol{\Sigma}_2^H$	$\boldsymbol{\Sigma}_3^R$	.17	0.97	.95	.11	1.08	.96	.08	1.12	.97
$\boldsymbol{\Sigma}_3^*$	$\boldsymbol{\Sigma}_2^R$	$\boldsymbol{\Sigma}_3^R$	.18	0.98	.95	.11	1.09	.97	.08	1.14	.97
$\boldsymbol{\Sigma}_3^{**}$	$\boldsymbol{\Sigma}_2^R$	$\boldsymbol{\Sigma}_3^R$	.18	1.00	.95	.11	1.10	.97	.08	1.12	.97

Note: $\boldsymbol{\Sigma}_3^*$ is the 1^{st} and $\boldsymbol{\Sigma}_3^{**}$ the 2^{nd} order correction, as defined in equation 17 and 18, and $\boldsymbol{\Sigma}^H$ and $\boldsymbol{\Sigma}^R$ are the Hessian based and robust estimators

Table 3.6: Fit of different latent class models to five indicators of tolerance for nonconformity from the 1976/77 General Social Survey (N=2689). Shown are the likelihood ratio (L^2), degrees of freedom (df), Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), and the largest bivariate residual (BVR) in the pairwise cross-table of the five indicators.

	L^2	df	p	BIC	AIC	max(BVR)
Independence model	4695.8	26	0.00	4490.5	4643.8	1048.5
2-class model	240.8	20	0.00	82.9	200.8	13.2
3-class model	48.7	14	0.00	-61.7	20.7	6.0
4-class model	6.5	8	0.59	-56.6	-9.4	0.11
5-class model	2.9	2	0.24	-12.9	-1.1	0.08

Table 3.7: Parameter estimates (standard errors) for the four-class model: class sizes and conditional probabilities to give all-tolerant answers given the latent class.

	"Intolerant"		"Tolerant"		"Intolerant of right"		"Intolerant of left"	
Class size	.56	(0.02)	.23	(0.01)	.11	(0.03)	.10	(0.03)
Tolerance for...								
Atheists	.03	(0.01)	.98	(0.01)	.41	(0.06)	.61	(0.07)
Communists	.04	(0.01)	.95	(0.02)	.59	(0.11)	.27	(0.07)
Militarists	.05	(0.01)	.92	(0.02)	.34	(0.05)	.38	(0.06)
Racists	.08	(0.01)	.90	(0.02)	.02	(0.06)	.81	(0.20)
Homosexuals	.13	(0.01)	.96	(0.01)	.72	(0.07)	.56	(0.06)

Chapter 4

Robustness of stepwise latent class modeling with continuous distal outcomes

Abstract

Recently, several bias-adjusted stepwise approaches to latent class modeling with continuous distal outcomes have been proposed in literature and implemented in generally available software for latent class analysis. In this paper, we investigate the robustness of these methods to violations of underlying model assumptions by means of a simulation study. While each of the four investigated methods yield unbiased estimates of the class-specific means of distal outcomes when the underlying assumptions hold, three of the methods may fail to different degrees when assumptions are violated. Based on our study, we provide recommendations on which method to use under what circumstances. The differences between the various stepwise latent class approaches are illustrated by means of a real data application on outcomes related to recidivism for clusters of juvenile offenders.

This chapter is accepted for publication as Bakk, Z., & Vermunt, J. K. (forthcoming). Robustness of stepwise latent class modeling with continuous distal outcomes *Structural Equation Modeling*

4.1 Introduction

Latent class (LC) analysis is a method widely used in the social and behavioral sciences to group individuals based on their responses on a set of observed variables (Goodman, 1974). Examples include the creation of a clustering concerning tolerance toward nonconformity (McCutcheon, 1985) or a typology of psychological contract types (De Cuyper, Rigotti, Witte, & Mohr, 2008). Often, in LC analysis applications the interest lies not only in obtaining a clustering, but also in determining whether the classes differ with respect to one or more, possibly continuous, distal outcome variable. For example, De Cuyper et al. (2008) tested whether the means of variables related to well being differed for latent classes representing types of psychological contracts, Pastor, Barron, Miller, and Davis (2007) studied differences in academic achievement of college students across goal orientation clusters, and Mulder et al. (2012) compared the means of 70 variables measuring different aspects of recidivism across clusters of juvenile offenders.

The class-specific means of a distal outcome can be determined by either a one-step or a stepwise approach. In the one-step approach, the distal outcome is incorporated in the LC model as an additional response variable and the resulting expanded model is estimated in the usual way. But this approach has several disadvantages. The main problem is related to the fact that rather strong assumptions need to be made about the within-class distribution of the distal outcome, and if these assumptions are violated the original LC model may be completely distorted. A related issue is that it is problematic to deal with multiple distal outcomes, which may either be dealt with simultaneously, requiring strong assumptions about their joint distribution, or one by one, implying that the LC solution may change per distal outcome. Furthermore, since the interest lies in explaining differences across classes in the distal outcome, using the distal outcome as one of the variables defining the latent classes creates an unintended circularity. Because of these problems, researchers often prefer using a three-step approach in which one first builds the LC model without the distal outcome(s), then determines the class memberships, and subsequently investigate the relationship between class memberships and the distal outcome(s), say using a simple ANOVA (Bakk et al., 2013). However, a well known disadvantage of this approach is that the estimates obtained in the third step are attenuated because of the classification error introduced when assigning individuals to classes (Bolck et al., 2004).

Recently, alternative three-step approaches have been proposed which yield unbiased estimates of the class differences in the distal outcome (Bakk et al., 2013). One method, called ML involves estimating the class-specific means and variances by maximum likelihood while correcting for the classification errors (Vermunt, 2010; Bakk et al., 2013). Another approach based on the work of Bolck et al. (2004), which we therefore call BCH approach, involves performing a weighted ANOVA, with weights which are inversely related to the classification error probabilities (Vermunt, 2010; Bakk et al., 2013). Both approaches can be used with either equal or unequal variance across classes.

A different type of stepwise approach for dealing with distal outcomes was proposed by Lanza et al. (2013). This approach, which we will refer to as the LTB approach, involves estimating a LC model in which the distal outcome of interest is used as a covariate predicting class membership using a logistic model rather than as a response variable

affected by the classes. As a second step, the class-specific means for the distal outcome are calculated from the parameters of the estimated LC model using Bayes theorem.

While simulation studies have shown that these three recently developed stepwise approaches (ML, BCH, and LTB) yield unbiased estimates of the class-specific means of distal outcomes when all underlying model assumptions hold (Bakk et al., 2013; Lanza et al., 2013), it is unknown whether these methods are robust for violations of these assumptions. For example, the ML and BCH approaches assume that the distal outcome is normally distributed within classes. While the ML approach is expected to be affected by violations of this assumption (Asparouhov & Muthén, 2014), the BCH approach is probably more robust since it is similar to a standard ANOVA. At the same time, for continuous variables, the LTB approach assumes that the relationship between the latent classes and the distal outcome is linear on a logit scale, and it is unknown whether violation of this assumption will bias the estimates of the class-specific means.

In the remaining of this paper, we first introduce the different types of stepwise approaches and describe their assumptions. Subsequently, in a simulation study, we compare the performance of the various approaches when certain underlying assumptions are violated. Next, we illustrate the methods with an analysis of a data example on juvenile recidivism. We end with a discussion and recommendations regarding the use of the different methods.

4.2 The basic LC model and extensions

In the following, we first introduce the basic LC model. Then, we describe different ways to deal with distal outcomes; that is, the simultaneous or one-step method, the LTB approach, and the three-step ML and BCH approaches. Special attention is dedicated to the assumptions made by the various approaches.

4.2.1 The basic LC model

Let Y_{ik} denote the response of individual i on one of K categorical response variables, where $1 \leq k \leq K$ and $1 \leq i \leq N$. The full response vector is denoted by $\mathbf{Y}_i$. LC analysis assumes that individuals belong to one of the T categories of an underlying categorical latent variable X which affects the responses (McCutcheon, 1987; Goodman, 1974; Hagenaars, 1990). Denoting a particular latent class by t, the model can be formulated as follows:

$$P(\mathbf{Y}_i) = \sum_{t=1}^{T} P(X = t) P(\mathbf{Y}_i | X = t), \tag{4.1}$$

where $P(X = t)$ represents the (unconditional) probability of belonging to class t and $P(\mathbf{Y}_i|X = t)$ represents the class-specific distribution of the responses $\mathbf{Y}_i$. These class-specific distributions are simplified further by assuming that the K response variables are independent within classes, which is known as the local independence assumption. This

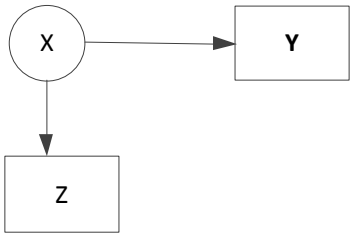

Figure 4.1: One-step approach

yields:

$$P(\mathbf{Y}_i) = \sum_{t=1}^{T} P(X = t) \prod_{k=1}^{K} P(Y_{ik}|X = t). \tag{4.2}$$

For categorical responses, $P(Y_{ik}|X = t) = \prod_{r=1}^{R_k} \pi_{ktr}^{I(Y_{ik}=r)}$, where π_{ktr} is the probability of giving response r on variable k for class t, and $I(Y_{ik} = r)$ is an indicator variable taking on the value 1 if $Y_{ik} = r$ and 0 otherwise.

The basic LC model described in Equation 4.1 can be extended to include a continuous distal outcome denoted by Z_i (visualized in Figure 4.1). This yields the following joint model for $\mathbf{Y}_i$ and Z_i:

$$P(\mathbf{Y}_i, Z_i) = \sum_{t=1}^{T} P(X = t) \prod_{k=1}^{K} P(Y_{ik}|X = t) f(Z_i|X = t), \tag{4.3}$$

where $f(Z_i|X = t)$ denotes the class-specific distribution of Z_i, which for continuous distal outcomes is typically defined to be a normal distribution with mean μ_t and variance σ_t^2. Note that the distal outcome serves as an additional response variable in the LC model.

The main disadvantage of this simultaneous modeling procedure is that the inclusion of Z in the model can alter the meaning of the classes (Petras & Masyn, 2010), especially when the normal distribution assumption for $f(Z_i|X = t)$ does not hold. Such a misspecification can even lead to over-extraction of the classes (Bauer & Curran, 2003). Moreover, when re-specifying the model for a different outcome variable, the definition of the classes may change. Another disadvantage is that from a substantive perspective it is undesired that the distal outcome contributes to the definition of the classes, that is, it creates a kind of circularity. To prevent these problems, alternative methods were proposed that we present in the following.

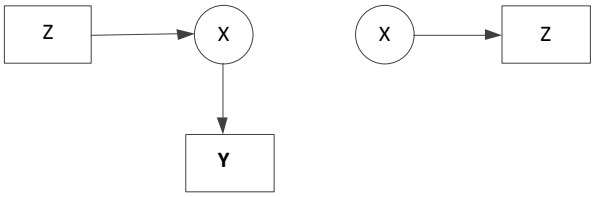

Figure 4.2: The two steps of the LTB approach

4.2.2 The LTB approach

To overcome the problems of the one-step approach resulting from the normal distribution assumption for Z, Lanza et al. (2013) proposed an alternative procedure which does not require making such an assumption. The LTB approach is a two-step procedure, which proceeds as follows:

Estimate a LC model in which Z is included as a covariate instead of a response variable (Figure 4.2.1).

Based on the estimates from the first step, calculate the class-specific means for Z (Figure 4.2.2).

In the first step, "covariate" Z is added to the model by extending the basic LC model described in Equation 4.1 as follows:

$$P(\mathbf{Y}_i|Z_i) = \sum_{t=1}^{T} P(X = t|Z_i) \prod_{k=1}^{K} P(Y_{ik}|X = t), \tag{4.4}$$

where $P(X = t|Z_i)$ denotes the probability of belong to class t given the "covariate" value Z_i. This probability is modeled by a multinomial logistic regression equation:

$$P(X = t|Z_i) = \frac{e^{\alpha_t + \beta_t Z_i}}{\sum_{s=1}^{T} e^{\alpha_s + \beta_s Z_i}}, \tag{4.5}$$

with intercepts α_t and slopes β_t.

The second step involves computing the class-specific means μ_t. It should be noted that these can be obtained as follows:

$$\mu_t = \int_Z Z \; f(Z|X = t), \tag{4.6}$$

where $f(Z|X = t)$, the class-specific distribution of Z, can be calculated using Bayes theorem as follows (Lanza et al., 2013):

$$f(Z|X=t) = \frac{f(Z)P(X=t|Z)}{P(X=t)}. \tag{4.7}$$

The quantities $P(X=t|Z)$ and $P(X=t)$ can be obtained from the estimated LC model, but $f(Z)$ is unknown. Lanza et al. (2013) suggested approximating $f(Z)$ by a kernel density estimate, and calculate the class-specific mean of Z using this estimate. However, as suggested by Asparouhov and Muthén (2014), a much simpler solution is to use the empirical distribution of Z, which involves replacing the integral in Equation 4.6 by a sum over the N sample units and replacing $f(Z)$ in Equation 4.7 by $\frac{1}{N}$. This yields:

$$\mu_t = \sum_{i=1}^{N} Z_i \frac{P(X=t|Z_i)}{N\ P(X=t)}. \tag{4.8}$$

This is how the LTB method is implemented in Mplus7.1 (Muthén & Muthén, 1998-2012) and LatentGOLD 5.0 (Vermunt & Magidson, 2013).

Lanza et al. (2013) did not discuss standard error estimation for the μ_t, implying that they did not solve the statistical testing problem. However, Asparouhov and Muthén (2014) suggested estimating these SEs as the square root of the within-class variance divided by the class-specific sample size; that is, as $\sigma_t^2/[N\ P(X=t)]$, where

$$\sigma_t^2 = \sum_{i=1}^{N} (Z_i - \mu_t)^2 \frac{P(X=t|Z_i)}{N\ P(X=t)}. \tag{4.9}$$

These SE estimates seem to somewhat underestimate the actual variation (Asparouhov & Muthén, 2014), which is probably caused by the fact that the uncertainty about the individuals' class memberships is not accounted for.

The simulation study by Lanza et al. (2013) showed that when generating Z from normal distribution with different means but equal variances (homoskedastic errors), the LTB estimates of the class-specific means are unbiased. It should be noted that in this situation the relationship between Z and X is linear-logistic, which is a well-known result on the relationship between linear discriminant analysis and logistic regression analysis (Agresti, 2002, p. 335). In other words, Lanza et al. (2013) looked only at the situation in which the "covariate" model is correctly specified. However, in other situations the relationship between Z and X may not be linear-logistic, in which case applying the LTB method may yield biased estimates of the class-specific means. This occurs, for example, when Z is normally distributed but with unequal variances across classes (when errors are heteroskedastic). In our simulation study, we investigate whether violating the linear-logistic association assumption of the "covariate" part of the LC model leads to biased estimates of the class-specific means.

A limitation of the LTB approach is that it cannot be used with multiple distal outcomes. A possible way out is to repeat the LTB analysis for every distal outcome, but in doing so there is no guarantee that the LC solution will remain the same across analyses. Moreover, when there are missing values on the Z variables, also the sample may vary per distal outcome, which may yield additional differences in the definition of the latent classes. As the one-step approach, the LTB is also affected by the fact that the classes are partially defined by Z, the outcome variable one wishes to predict.

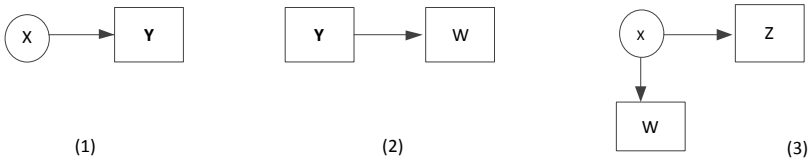

Figure 4.3: The three steps of the bias-adjusted three-step approaches

4.2.3 The bias-adjusted three-step approaches

We will now discuss the bias-adjusted three-step approaches for dealing with continuous distal outcomes. These proceeds as follows:

1. Build a standard LC model based on the categorical response variables (Figure 4.3.1).
2. Assign individuals to latent classes (Figure 4.3.2). The assigned class memberships are denoted by W.
3. Estimate the association between X and Z using the assigned class memberships W, taking into account that these contain classification errors (Figure 4.3.3).

In the first step, a model is built for response variables $\mathbf{Y}_i$ using the basic LC model described in Equation 4.1 and depicted in Figure 4.3.1. In the second step, individuals are assigned to latent classes based on their posterior class membership probabilities Z (Figure 4.3.2). These are calculated from the parameters of the step-one model using Bayes rule:

$$P(X = t|\mathbf{Y}_i) = \frac{P(X = t)P(\mathbf{Y}_i|X = t)}{P(\mathbf{Y}_i)}. \tag{4.10}$$

Two possible assignment rules are modal and proportional assignment, which in cluster analysis terminology yield a hard and a soft partitioning, respectively (Dias & Vermunt, 2008). In modal assignment, each individual is assigned to a single class; that is, the class for which the posterior membership probability is largest. This can be expressed as follows: $P(W = s|\mathbf{Y}_i) = 1$ if $P(X = s|\mathbf{Y}_i) > P(X = t|\mathbf{Y}_i)$ for all $s \neq t$ and equals to 0 for the other classes. In proportional assignment, each individual is assigned to each of the classes with a weight equal to $P(W = s|\mathbf{Y}_i) = P(X = s|\mathbf{Y}_i)$. In practice, this implies that subsequent analyses should be performed using an expanded data with T records for each unit, with records weights equal to $P(W = s|\mathbf{Y}_i)$.

Irrespective of the assignment method used, classification errors will be present unless the classification is perfect (Bolck et al., 2004). By aggregating over the observed data patterns, the amount of errors can be expressed as the probability of an assigned class membership s conditional on the true class membership t (Vermunt, 2010; Bakk et al.,

2013),

$$P(W = s|X = t) = \frac{\sum_{i=1}^{N} P(X = t|\mathbf{Y}_i)P(W = s|\mathbf{Y}_i)}{N\ P(X = t)}. \tag{4.11}$$

Lastly, in the third step, the class assignments W are used to estimate the relation between X and Z while correcting for the known classification errors introduced in step two (Figure 4.3.3). This is achieved using a model of the form (Bakk et al., 2013):

$$P(W = s, Z_i) = \sum_{t=1}^{T} P(X = t)f(Z_i|X = t)P(W = s|X = t). \tag{4.12}$$

Note that this is a LC model in which Z and W are used as response variables, and in which $P(W = s|X = t)$ is fixed. The model described in Equation 4.12 can be either estimated using maximum likelihood estimation (yielding the ML approach) or using a weighted analysis as proposed by Bolck, Croon and Hagenaars (yielding the BCH approach) (Bolck et al., 2004; Vermunt, 2010). These two approaches are presented in more detail in the following.

The ML approach

The ML approach estimates the LC model defined in Equation 4.12 directly. The $P(W = s|X = t)$ are fixed to their estimated values from the second step (see Equation 4.11), while the parameters in the part of interest, $f(Z_i|X = t)$, are freely estimated. To be able to estimate the class-specific means μ_t, we need to specify the distributional form of $f(Z_i|X = t)$, which for continuous Z is usually defined to be a normal distribution. The variance of Z can be modeled as either equal or unequal across classes, which we refer to as the ML(equal) and ML(unequal) approaches. Standard error estimates for the free parameters are obtained based on the robust estimator, which is especially needed when proportional assignment is used (Bakk, Oberski, & Vermunt, 2014). Tests for the equality of means can be performed using Wald tests.

The ML approach yields unbiased estimates of the μ_t and their standard errors when normality assumption holds (Bakk et al., 2013, 2014). However, the approach may fail when this assumption is violated. For example, when Z has a bimodal distribution within the classes, the step-three LC model may pick up this bi-modality in Z which will fully distort the original definition of the latent classes (Asparouhov & Muthén, 2014). To decrease the likelihood of obtaining a completely different LC solution, both in the Mplus 7.1 and the LatentGOLD 5.0 implementation specific starting values are used for the step-three model.[2] In this way, a local maximum of the likelihood is obtained with class definitions which are closer to those of first-step model than of the global maximum (Asparouhov & Muthén, 2014).[3]

[2]Note that the ML approach is called 3step in Mplus 7.1 and is an option of 'Auxiliary'

[3]Mplus 7.1 estimates the step-three model using as starting values the estimated class sizes from the first step, while Latent GOLD 5.0 fixes these values. For the means of Z, Mplus 7.1 uses the unadjusted class-specific means, while Latent GOLD 5.0 starts using the overall mean and variance of Z for all classes. Due to this different implementation in some cases different results can be obtained, though this is rare and occurs mainly in situations where the use of the ML approach is anyway not recommended.

The ML approach may also lead to biased estimates for μ_t if the error variance is wrongly assumed to be equal across the classes. However, it is less clear how problematic such a misspecification will be. In the simulation study, we investigate the impact of bimodality and of wrongly assuming homoskedastic variances when they are heteroskedastic.

The BCH approach

While the ML approach estimates the LC model defined in Equation 4.12 directly, the BCH approach transforms the problem and estimates an ANOVA model with observed variables only. It "recreates" the true latent classification by weighting W with the inverse of the classification errors (Bolck et al., 2004, Vermunt, 2010). The resulting model is estimated using a pseudo maximum likelihood estimation procedure. To account for the multiple (T) records per individual and for the weighting, robust standard errors should be used (Vermunt, 2010, Bakk et al., 2014). The equality of class-specific means is tested using Wald tests.

An important advantage of the BCH approach compared to the ML approach is that the class definitions will not change when the distribution of Z is misspecified. The reason for this is that it involves performing an ANOVA-like analysis with observed variables only rather than estimating a LC model. Moreover, a positive side effect of using robust standard errors is that these correct for all kinds of misspecifications, thus also for a possible misspecification of the distribution of the errors. This means, for example, that Wald tests for the class-specific means are identical irrespective of whether one assumes homoskedastic or heteroskedastic errors. So, we can simply assume equal error variances when using the BCH approach.

A possible problem associated with the BCH approach, which may occur with very low class separation and small sample sizes, is that the error variance may become negative in one or more classes when these variances are specified to be unequal across classes. In these situations, it is recommended not to use the BCH approach with class-specific variances. A similar problem was reported by Bakk et al. (2013) for nominal outcomes, where negative cell frequencies could be found. However, negative variances will not occur when using the BCH method with equal variances, which is therefore the preferred approach.

The BCH approach is available in Latent GOLD 5.0, with the default specification for continuous distal outcomes being the one with equal variances. The Mplus 7.1 version available at the time we performed this research did not implement the BCH approach, but we were notified that it will become available in a next version.

4.2.4 A comparison of the underlying assumptions

Table 4.1 summarizes the assumptions of the various approaches and the possible consequences of their violations. As can be seen, for certain assumptions it is known that their violation will have an impact on the estimated class-specific means and/or their SEs. For example, violation of the normality assumption, such as when class-specific distributions are bimodal, may bias the results obtained with the ML method (Asparouhov & Muthén,

Table 4.1: Underlying assumptions of the stepwise LC analysis approaches and hypothesized consequences of their violation

	BCH	ML(equal)	ML(unequal)	LTB
Assumptions:				
Normal distribution	yes	yes	yes	no
Linear-logistic Z-X relationship	no	no	no	yes
Equal variances	yes	yes	no	yes
No uncertainty in step-one parameters	yes	yes	yes	yes
Violation hypothesized to affect:				
Normal distribution	none	means, SEs	means, SEs	–
Linear-logistic Z-X relationship	–	–	–	means
Equal variances	none	means, SEs	–	means
No uncertainty in step-one parameters	SEs	SEs	SEs	SEs

2014). It is also known that plugging in the parameter values from step one without accounting for their sampling fluctuation can lead to underestimated SEs when using the three-step approaches (Bakk et al., 2014). The same seems to apply to the somewhat ad hoc SE estimates proposed by Asparouhov and Muthén (2014) for the LTB approach. However, the extent of these effects have not systematically been investigated so far. Moreover, for some of the assumptions it is unknown whether their violation will cause bias. For instance, we do not know how strongly the LTB method is affected by a possible logistic non-linearity of the true Z-X association.

4.3 Simulation study

The goal of the simulation study is to compare the different stepwise LC methods for dealing with continuous distal outcomes with regard to their robustness against violations of underlying assumptions. We will focus on the assumptions summarized in Table 4.1. More specifically, we generated data under different degrees of bimodality and heteroskedasticity. Bimodality violates the assumption of normality made by the BCH and the two ML approaches. Heteroskedasticity violates the assumption of equal variances of the ML(equal) approach and the assumption of logistic linearity of the LTB method. We not only investigate bias, but also the quality of the SE estimates provided by the various stepwise methods. The results of two studies are presented, in the first study we focus solely on bias, whereas in the second study the consequences of sampling fluctuation are also considered.

4.3.1 Study 1

In Study 1, we investigated the bias in the estimated class-specific means of the distal outcome for each of the stepwise methods. The population model was a 2-class model for 6 dichotomous response variables. Class sizes were set to be equal. The class-specific probability of a positive answer was set to .80 in class 1 for all indicators, and to .20 in

class 2, corresponding to an entropy-based R^2 value of .82 [4]. Both modal and proportional assignment were used when applying the three-step approaches.

Furthermore, we manipulated the degree of heteroskedasticity and bimodality in the class-specific distribution of Z. We defined four conditions with gradually increasing degrees of heteroskedasticity. We set the variance in class 1 equal to 1, and in class 2 equal to 1, 4, 9, and 25, respectively, thus going from equal to highly unequal variances. In addition, we defined three conditions with gradually increasing degrees of bimodality. In class 1, Z was assumed to come from the mixture of normal distributions of the form $0.75N(-2, \tau^2) + 0.25N(2, \tau^2)$, thus having a class-specific mean of -1. In class 2, Z followed the mixture distribution $0.75N(2, \tau^2) + 0.25N(-2, \tau^2)$, thus having a class-specific mean of 1. The degree of bimodality was manipulated by setting the variance σ^2 to either 0.01, 0.50, or 1, which affects the overlap between the mixing distributions. In this way, conditions were created without any overlap at all and thus extreme bimodality ($\tau^2 = 0.01$), some overlap and thus moderate bimodality ($\tau^2 = 0.50$), and large overlap and thus nonextreme bimodality ($\tau^2 = 1$).

In all investigated conditions, we generated a single data set with 1,000,000 observations. For each condition and each estimator, we determined the bias in the difference in means between the two classes (the true difference is 2). The ML estimator was applied with equal and unequal variances, and for both of these methods we used versions with random and with predefined starting values. Because results were very similar for modal and proportional assignment, we report only the results obtained with modal assignment.

Table 4.2 summarizes the results for the four heteroskedasticity conditions. As can be seen in the first row, when the variances are equal between the two classes, all the methods perform well. This result is similar to what was found in previous simulation studies (Bakk et al., 2013; Lanza et al., 2013; Asparouhov & Muthén, 2014). However, as the degrees of heteroskedasticity increases (variance in class 2 increases), the ML(equal) estimate becomes more strongly biased. The ML approach with random starting values shows larger bias than its counterpart with predefined starting values. At the same time, the BCH and the ML(unequal) approaches obtain unbiased estimates in all conditions. The estimates obtained with the LTB approach have a small bias when the variances are unequal, which shows that the method is indeed affected by the relationship between Z and X being nonlinear on the logit scale.

Table 4.3 summarizes the conditions where Z has a bimodal within-class distribution. In all three conditions, the BCH and LTB methods obtain the correct class-specific means. In contrast, the ML approaches are highly sensitive to bimodality. When the bimodality becomes more extreme, the bias of the ML estimates increases. The ML approaches with prespecified starting values show somewhat smaller bias than their counterparts with the random starting values.

In summary, the LTB and the ML(equal) approaches yielded biased estimates when variances are unequal between classes. Moreover, both ML approaches were sensitive for bimodality. The ML approaches with predefined starting values (as implemented as default setting in Mplus 7.1 and LatentGOLD 5.0) performed somewhat better than

[4] Note that in both the data generating and step-one models (for the three-step methods) the response probabilities were constrained across the items to be equal, however conclusions remain unchanged with the unconstrained model. The same constraint was active in Study 2 as well

Table 4.2: Absolute bias in the estimated difference of means obtained with the stepwise LC methods for varying degrees of heteroskedasticity using modal assignment in Study 1

Heteroskedasticity	BCH	ML(equal, nonrandom)	ML(unequal, nonrandom)	ML(equal, random)	ML(unequal, random)	LTB
none	0.00	0.00	0.00	0.00	0.00	0.00
low	0.00	0.15	0.00	0.19	0.00	0.04
medium	0.01	0.05	0.01	0.06	0.01	0.04
high	0.00	0.10	0.00	0.11	0.00	0.03

Table 4.3: Absolute bias in the estimated difference of means obtained with the stepwise LC methods for varying degrees of bimodality using modal assignment in Study 1

Bimodality	BCH	ML(equal, nonrandom)	ML(unequal, nonrandom)	ML(equal, random)	ML(unequal, random)	LTB
low	0.00	0.11	0.11	0.11	0.11	0.01
medium	0.00	0.12	0.12	0.12	0.12	0.01
high	0.01	0.21	0.21	2.00	2.00	0.01

their counterparts with random starting values. The BCH approach performed well in all conditions.

4.3.2 Study 2

In the following, we compare the different methods using a larger and more realistic LC model and, moreover, accounting for sampling fluctuation. Given that the ML methods with random starting values proved to be more biased than their counterparts that use predefined starting values, we restrict ourselves to the later implementation.

Data was generated from a 4-class model with 8 dichotomous indicators, with parameter values based on the application to psychological contract types described by Bakk et al. (2013). The class sizes were set to .50, .30, .10, and .10, similarly to this real data example used as starting point. In class 1, all indicators have a high probability of a positive answer, while in class 2 the first 4 indicators have a high probability of a positive answer, and the last 4 of a negative answer. At the same time, in class 3 the first 4 indicators

Table 4.4: Bias and coverage rate for one class-specific mean for different degrees of heteroskedasticity and bimodality

	BCH	ML	LTB	BCH	ML	LTB
heteroskedsticity:		None			High	
Bias	0.00	0.00	0.00	-0.01	-0.24	-0.50
Coverage	0.94	0.93	0.88	0.96	0.86	0.64
bimodality:		Low			High	
Bias	0.00	-0.05	-0.02	0.00	-0.50	-0.01
Coverage	0.94	0.87	0.84	0.93	0.01	0.84

have a low probability of a positive answer, while the last 4 have a high probability of a positive answer, and in class 4 all indicators have a low probability of a positive answer. We manipulated the class separation by setting the probability of a positive answer to .80 (.20) or .90 (.10), which yields a moderate and a high separation condition.

The class-specific means of the distal outcome were set to -1, -0.5, 0.5, and 1, respectively. Similarly to Study 1, we looked into two types of situations: unequal class-specific variances of Z (heteroskedasticity) and bimodal class-specific distributions of Z. More specifically, the conditions with heteroskedasticity were created by setting the variance of Z to 1, 4, 9, and 25 in class 2 and 3, while keeping it equal to 1 in class 1 and 4. The bimodal conditions were defined such that class 2 and 3 have bimodal distributions, while class 1 and 4 have unimodal distributions. The bimodality was again obtained by using the mixture distribution: $0.75N(-1, \tau^2) + 0.25N(1, \tau^2)$ in class 2 and $0.75N(1, \tau^2) + 0.25N(-1, \tau^2)$ in class 3. We manipulated the extremeness of the bimodal distributions by setting τ^2 equal to 1, 0.5, and 0.01.

Three sample size conditions were investigated: 500, 1000, 2000. For all conditions, 500 replications were used. The bias in class-specific means and the coverage rate based on the SE estimates was investigated for the LTB, BCH, ML(equal), and ML(unequal) approach, where by coverage we mean the proportion of the time that the interval contains the true value of interest. We consider an estimator to perform well if it has a bias close to 0 and a coverage rate close to .95.

Results under heteroskedasticity

Table 4.5 shows the results averaged across separation level and sample size conditions for all levels of heteroskedasticity. Note that the 4 different levels of heteroskedasticity (none, small, medium, and large) correspond with a variance of 1, 4, 9, and 25 in class 2 and 3. Similarly to Study 1, the estimates obtained with the BCH approach are unbiased in all conditions. Moreover, coverage rates are between .93 and .95, thus slightly too low in some conditions. Also the ML approaches yield results comparable to Study 1. That is, in all conditions the estimates obtained with the ML(unequal) approach are unbiased. Coverage rates are between .93 and .95. With ML(equal) the estimates are highly biased.

At the same time, the LTB estimates are increasingly biased as the degree of heteroskedasticity increases. While the first class is hardly biased, the other classes are strongly affected, especially class 2. The coverage rates obtained with LTB are too low even in the none heteroskedastic condition (between .86 and .91), although the estimated class-specific means are unbiased. This shows that the undercoverage is the result of an underestimation of the SEs.

Table 4.6 presents the bias and coverage rate for the mean of class 2, but now separately for each separation level and sample size condition. We chose to give the detailed results only for the mean of class 2 because this parameter showed the largest bias. For all methods at hand, it can be seen that as uncertainty increases (smaller sample size and lower separation) the bias increases and the coverage rate decreases. Only the BCH approach obtains almost unbiased estimates and good coverage rates in all conditions. However, even this method obtains a somewhat too low coverage rate with low separation and small sample size (between .91 and .93), thus showing that the SEs are somewhat

Table 4.5: Bias and coverage rate for class-specific means averaged across separation level and sample size conditions for different degrees of heteroskedasticity. Study 2

		Bias				Coverage			
Method	Heter.	$X=1$	$X=2$	$X=3$	$X=4$	$X=1$	$X=2$	$X=3$	$X=4$
BCH	none	0.00	0.00	0.01	-0.01	.94	.94	.94	.93
	low	0.00	0.00	0.01	0.01	.94	.94	.94	.94
	medium	0.00	0.00	0.00	0.00	.94	.94	.95	.94
	high	0.00	-0.01	0.01	0.00	.94	.96	.94	.94
ML(eq.)	none	0.00	0.00	0.00	-0.01	.94	.93	.93	.93
	low	0.01	-0.10	0.01	0.17	.94	.84	.91	.77
	medium	0.02	-0.18	-0.05	0.36	.93	.81	.90	.70
	high	0.03	-0.24	-0.12	0.53	.93	.86	.89	.77
ML(uneq.)	none	0.00	0.00	0.00	-0.01	.94	.93	.92	.93
	low	0.00	0.00	0.00	-0.01	.94	.94	.93	.94
	medium	0.00	0.00	0.00	0.00	.95	.94	.94	.94
	high	0.00	-0.01	-0.01	0.00	.94	.95	.94	.95
LTB	none	0.00	0.00	0.00	0.00	.91	.88	.86	.86
	low	0.00	-0.18	0.11	0.00	.97	.65	.71	.94
	medium	-0.02	-0.48	0.22	0.12	.94	.58	.64	.83
	high	-0.03	-0.50	0.21	0.29	.91	.64	.64	.86

Note. heter=heteroskedasticity, eq= equal

underestimated in these situations.

It can be seen that the LTB method is very sensitive to the stability of the step-one model: in the low separation conditions the bias is extremely large, while in the high separation conditions the bias is negligible. The coverage rates with the LTB approach are clearly too low (between .87 and .94), even in the conditions with equal variances, which shows that the SEs are underestimated. At the same time, the ML methods are less affected by the class separation. If the variance of Z is correctly specified, the ML estimates are unbiased in all conditions, obtaining a coverage rate between .92 and .94 (thus having a minor undercoverage). However, if the variances are wrongly assumed to be equal, the bias is always large.

Results with bimodality

The results for the three bimodality conditions which are presented in Table 4.7. These are again averages across sample size and separation level conditions. The BCH estimates are unbiased in all conditions and their coverage rates are between .90 and .94, with lowest coverage rates occurring in the most extreme bimodality condition.

At the same time, the ML(equal) approach fails when the bimodality is the most extreme, but as the bimodality becomes less extreme the bias decreases. The bias in the ML(unequal) estimates is lower than in those of ML(equal). However, even the ML(unequal) estimates are much worse than the LTB and BCH estimates. Both ML approaches show much too low coverage rates, but these are in fact uninformative given that these estimates are biased anyhow. Furthermore, the LTB approach yields estimates with very small bias, however, the coverage rate is again too low (.77 at its worst), especially in class 2 and 3 which have a bimodally distributed Z.

Table 4.6: Bias and coverage rate for the mean of class 2 per separation level and sample size condition for different degrees of heteroskedasticity. Study 2

Method	Heter.	Low separation						High separation					
		N=500		N=1000		N=2000		N=500		N=1000		N=2000	
		Bias	Cov.	Bias	Cov.	Bias	Cov.	Bias	Cov.	Bias	Cov.	Bias	Cov.
BCH	none	-0.01	.91	-0.01	.93	0.00	.93	0.00	.95	0.00	.93	0.00	.95
	low	-0.01	.95	0.00	.93	0.00	.93	0.00	.95	0.01	.94	0.00	.94
	medium	-0.01	.93	-0.01	.95	-0.01	.93	0.00	.95	0.01	.96	-0.01	.95
	high	-0.02	.94	-0.01	.95	0.00	.96	0.00	.96	-0.01	.97	-0.01	.96
ML(eq.)	none	-0.01	.92	0.00	.92	0.00	.95	0.00	.94	0.00	.93	0.00	.95
	low	-0.15	.85	-0.15	.76	-0.16	.64	-0.05	.94	-0.04	.93	-0.05	.91
	medium	-0.27	.80	-0.29	.72	-0.31	.54	-0.07	.94	-0.07	.95	-0.09	.89
	high	-0.40	.84	-0.40	.77	-0.42	.68	-0.07	.96	-0.08	.97	-0.08	.93
ML(uneq.)	none	0.00	.92	0.00	.92	0.00	.93	0.00	.95	0.00	.93	0.00	.95
	low	-0.01	.94	0.00	.93	0.00	.94	0.00	.95	0.01	.94	0.00	.94
	medium	0.00	.94	-0.02	.94	-0.01	.94	0.00	.94	0.01	.96	-0.01	.95
	high	-0.03	.93	-0.01	.94	-0.01	.94	0.00	.95	-0.02	.96	-0.01	.96
LTB	none	0.00	.87	0.00	.87	0.00	.90	0.00	.94	0.00	.94	0.00	.92
	low	-0.44	.49	-0.37	.49	-0.25	.50	-0.01	.82	0.01	.81	0.00	.80
	medium	-1.05	.35	-0.88	.42	-0.90	.39	-0.02	.76	0.00	.81	-0.02	.76
	high	-1.17	.47	-0.95	.49	-0.81	.55	-0.01	.79	-0.03	.79	-0.02	.74

Note. heter=heteroskedasticity, eq= equal, Cov=coverage

Table 4.7: Bias and coverage rate for the class-specific means averaged across separation level and sample size conditions for different degrees of bimodality. Study 2

		Bias				Coverage			
Method	Bimodality	$X=1$	$X=2$	$X=3$	$X=4$	$X=1$	$X=2$	$X=3$	$X=4$
BCH	low	0.00	0.00	0.01	-0.01	.93	.94	.93	.93
	medium	0.00	0.00	0.00	-0.01	.93	.94	.93	.93
	high	0.00	0.00	0.01	0.00	.90	.93	.93	.92
ML(eq.)	low	0.00	-0.05	0.04	0.07	.94	.87	.90	.88
	medium	0.00	-0.13	0.16	0.10	.93	.62	.69	.68
	high	0.00	-0.50	0.49	0.00	.95	.01	.01	.95
ML(un.)	low	-0.01	0.00	0.02	0.01	.93	.94	.91	.93
	medium	-0.02	-0.02	0.05	0.04	.90	.92	.87	.87
	high	0.00	-0.17	0.26	0.01	.93	.65	.47	.90
LTB	low	0.00	-0.02	0.03	-0.01	.94	.84	.84	.94
	medium	0.01	-0.03	0.04	-0.02	.96	.77	.81	.95
	high	0.01	-0.01	0.03	-0.03	.95	.84	.82	.94

Table 4.8 presents the results for the mean in class 2, but now separately per sample size and separation condition. The BCH approach is unbiased. In the low separation conditions, it has a coverage rate between .89 and .93. However in the high separation condition the coverage rate is better.

The ML methods fail with the most extreme bimodality, which is also what we saw in Study 1. As the bimodality becomes less extreme, the estimates obtained using ML(unequal) become less biased. This tendency depends solely on the amount of bimodality and not on separation level or sample size. At the same time, using the LTB method, the bias is small in all conditions. However, in the low separation conditions, the coverage rate is too low irrespective of the sample size.

To conclude the ML approaches fail when Z has a bimodal distribution within classes. At the same time, in the heteroskedastic variance conditions, estimated class-specific means obtained with the ML approach are unbiased if the variance is modeled correctly. This shows that the ML methods are very sensitive to misspecification when dealing with continuous distal outcomes. The LTB approach turns out to yield biased estimates when the underlying assumption of logistic linearity is violated. In all conditions, the method that obtained the least biased estimates was the BCH approach. All investigated approaches yield too low coverage rates in the higher uncertainty conditions; that is, when class separation is lower and sample size is smaller.

4.4 Empirical example

To illustrate the stepwise LC modeling approaches we use a data set on juvenile offenders' recidivism from the Dutch Ministry of Justice, which was analyzed earlier by Mulder et al. (2012) using unadjusted three-step LC analysis. Though these authors were aware of the fact that this approach may yield biased estimates, for them it was the only way to proceed given the large number (70) of distal outcomes, which cannot be dealt with using a one-step approach (Mulder et al., 2012), and given that bias-adjusted approaches were not available at that time.

Table 4.8: Bias and coverage rate for the mean of class 2 per separation level and sample size condition for different degrees of bimodality. Study 2

Method	Bimodality	Low separation						High separation					
		N=500		N=1000		N=2000		N=500		N=1000		N=2000	
		Bias	Cov.	Bias	Cov.	Bias	Cov.	Bias	Cov.	Bias	Cov.	Bias	Cov.
BCH	low	0.00	.94	-0.01	.94	-0.01	.91	0.01	.95	-0.01	.94	-0.01	.97
	medium	0.00	.94	0.00	.93	0.00	.91	0.00	.95	0.00	.96	0.00	.95
	high	0.00	.89	0.00	.91	0.00	.91	0.00	.96	0.00	.95	0.00	.94
ML(eq.)	low	-0.07	.89	-0.08	.83	-0.09	.71	-0.02	.95	-0.03	.91	-0.03	.92
	medium	-0.18	.62	-0.20	.42	-0.20	.22	-0.07	.89	-0.07	.82	-0.07	.73
	high	-0.50	.00	-0.50	.00	-0.50	000	-0.48	.03	-0.49	.01	-0.50	.00
ML(uneq.)	low	0.00	.93	-0.01	.92	-0.01	.90	0.01	.96	-0.01	.94	-0.01	.96
	medium	-0.03	.91	-0.02	.90	-0.02	.87	-0.01	.96	-0.01	.96	0.00	.95
	high	-0.25	.42	-0.30	.38	-0.33	.35	-0.05	.92	-0.05	.92	-0.04	.91
LTB	low	-0.04	.78	-0.03	.78	-0.03	.76	0.00	.92	-0.01	.92	-0.01	.91
	medium	-0.08	.67	-0.06	.65	-0.05	.63	-0.01	.89	-0.01	.90	0.00	.89
	high	-0.04	.76	-0.02	.79	-0.02	.76	0.00	.92	-0.01	.91	0.00	.88

Table 4.9: Profile of latent classes of juvenile offenders

	Violent property	Property	Serious violent	Sexual
Class proportion	.46	.29	.15	.10
Number of offences				
low	.00	.32	.73	.80
medium	.30	.45	.27	.18
high	.70	.24	.00	.02
Misdemeanour	.56	.08	.20	.05
Drugs	.06	.03	.05	.00
Vandalism	.00	.00	.00	.00
Property	.99	1.0	.28	.17
Moderate violent	.91	.36	.74	.15
Violent property	.63	.82	.18	.01
Serious violent	.41	.05	.24	.00
Sexual same age	.14	.04	.11	.63
Pedosexual	.04	.00	.00	.61
Manslaughter	.08	.03	.43	.06
Arson	.05	.00	.10	.00
Murder	.01	.01	.18	.03

Table 4.10: Mean (and SE) of frequency and severity of recidivism for the four offender classes obtained with five stepwise LC approaches

	Violent property	Property	Serious violent	Sexual
Frequency of recidivism				
Unadjusted	9.85(.62)	6.08(.53)	3.33(.58)	2.78(.54)
BCH	10.46(.72)	5.73(.65)	3.03(.65)	2.74(.57)
ML(equal)	10.39(.65)	5.69(.48)	3.31(.49)	2.83(.51)
ML(unequal)	12.72(.76)	3.50(.51)	1.31(.47)	3.07(.61)
LTB	10.15(.66)	5.92(.50)	3.27(.40)	2.81(.43)
Severity of recidivism				
Unadjusted	5.60(.17)	4.80(.22)	3.75(.33)	2.13(.34)
BCH	5.73(.20)	4.77(.26)	3.75(.38)	2.02(.36)
ML(equal)	5.74(.20)	4.76(.27)	3.72(.39)	2.03(.35)
ML(unequal)	5.73(.20)	4.76(.27)	3.83(.40)	1.89(.50)
LTB	5.69(.17)	4.79(.22)	3.76(.30)	1.98(.31)

As Mulder et al. (2012), we built a LC model using 13 categorical indicators, representing the offense frequency prior to conviction (grouped into three categories: low, average, and high) and 12 types of offenses (yes/no). The model selected based on the BIC is the 4-class model (BIC = -3769), which turns out to be a rather strong clustering model in terms of class separation (Entropy-based R^2=.75). The four groups are, as shown in Table 4.9, the violent property offenders (being differentiated from the other groups by high scores on the property offenses and misdemeanor, and a high number of offenses), the property offenders (similar to first group, but lower offense rates), the serious violent offenders (with high scores on manslaughter and murder) and the sexual offenders (with high scores on sexual offenses with same age victims and pedophilic offenses).

While Mulder et al. (2012) built the step-one model using the full sample of 1082 respondents, we used a subsample of 728 respondents. This is the subsample for which recidivism information is available; that is, meeting the requirement of having been released to the community for a minimum of 2 years at the time of the data collection. We used this subsample instead of the full sample because the LTB approach requires that the distal outcomes are observed for all units. For comparability of results, the same sample was used for the three-step approaches as well, though for these approaches it is no problem to base the third step analysis on subsample of the sample used to build the LC model.[5]

We computed the class-specific means and their SEs of two distal outcome variables, the frequency and severity of recidivism, using the stepwise LC methods (see Table 4.10). The overall Wald test indicates that there is a significant difference between the class-specific means of the frequency of recidivism with all methods at hand. It can be seen, that as a result of the attenuation effect, the differences between the classes are somewhat smaller for the unadjusted 3-step approach. Irrespective of the method used, violent property offenders have the highest recidivism frequency, followed by the property offenders. All methods except ML(unequal) obtain a somewhat higher class-specific mean for the serious violent offenders than for sexual offenders, while this later method reverses the order of these two groups. This may be the result of the fact that ML(unequal) is more strongly affected by arbitrary deviations from within-class normality. The estimated class-specific means are similar for BCH and ML(equal), but somewhat different from LTB estimates. This may indicate that the linear-logistic assumption is violated to a certain extent. Note also that the SEs obtained with the LTB approach are smaller than those of the bias-adjusted three-step methods and sometimes even smaller than those of the unadjusted three-step analysis. This confirms that the LTB SEs are probably underestimating the actual sampling variability in the reported class-specific means.

Also for severity of recidivism, the overall Wald test shows a significant difference in means across classes for all methods at hand. Moreover, again the differences between the classes are somewhat smaller for the unadjusted 3-step approach. The severity is highest among violent property offenders, followed by property offenders, violent offenders, and sexual offenders. Note that all adjusted methods give almost the same class-specific means, showing that assumption violations are not a problem here. SE estimates are again probably too low for the LTB approach.

[5]The model parameters obtained with 1082 observation are very similar, which probably the result of the measurement model being strong.

4.5 Conclusions and discussion

We investigated the robustness of four stepwise LC analysis methods for studying the relationship between class membership and continuous distal outcomes. The BCH method, the ML method with equal variances, and the ML method with unequal variances are bias adjusted three-step approaches, which assume that the distal outcome is normally distributed, whereas ML(equal) also assumes homoskedasticity. The LTB method, which obtains the class-specific means of the distal outcome by estimating a LC model in which the distal outcome is treated as a covariate, assumes that the relationship between the distal outcome and class membership is linear on a logistic scale.

In a simulation study we investigated the performance of the stepwise methods under different degrees of heteroskedasticity and bimodality of the class-specific distributions of the distal outcome. Bimodality is a violation of the assumption of normality needed for the BCH and ML approaches; heteroskedasticity violates the assumption of logistic linearity of the LTB approach, and the assumption of homoskedasticity of the ML approach with equal variances. The simulation results revealed that the BCH method is the most robust approach: it yielded unbiased estimates under all investigated conditions. This is most probably the result of the fact that it involves performing a weighted ANOVA, a method that is known to be robust against violations of assumptions. The other methods are sensitive to the violations considered. The ML approach fails to different degrees in all the situations investigated. When the variance is heteroskedastic, modeling it as equal between the classes produces a bias in the class-specific means. However, if the heteroskedasticity is correctly modeled, the ML method works fine. The ML approach cannot handle bimodal class-specific distributions of the outcome variable and probably any other departure from normality as well.[6] The LTB approach yields biased estimates of the class-specific means when the errors are heteroskedastic, and shows a small bias with certain conditions of bimodality.

All four methods yielded coverage rates lower than the nominal .95 rate when the separation between classes is low and the sample size is small. For the three-step approaches, the too low coverage rate is caused by ignoring the uncertainty about the fixed parameter estimates from step one (Bakk et al., 2014). However, by taking this uncertainty into account, coverage rates close to the nominal .95 level can be obtained, as shown by Bakk et al. (2014) for the ML approach. For the LTB approach a somewhat ad hoc SE estimator was used, which turned out to yield a too low coverage in all investigated conditions, the undercoverage increases when the uncertainty about the step-one parameters and about the class memberships increases.

Because it performed very well in all investigated conditions, we recommend using the BCH approach for stepwise LC modeling with continuous distal outcomes. The use of the ML methods with continuous distal outcomes is recommended only with precaution due to its sensitivity to assumption violations. We also recommend caution with the LTB method, both due to the bias that can occur with heteroskedastic errors and due to the undercoverage resulting from the current SE estimates. The application to the juvenile recidivism data showed that results may indeed differ depending on the method that is

[6] Note that Mplus gives a warning when the definition of the classes changes due to deviations from normality. When this problem occurs, the BCH approach should clearly be preferred.

used. It seems to be safest to rely on the results obtained with the very robust BCH approach.

While in this paper we focused on the performance of stepwise LC analysis approaches for the simple case in which one studies the relationship between class membership and a single continuous distal outcome, it should be mentioned that these approaches can also be used in much more general situations. The ML and BCH three-step approaches are very flexible, and can, for instance, also be applied with covariates (Vermunt, 2010), with multiple latent variables (Bakk et al., 2013), with continuous indicators (Gudicha & Vermunt, 2011), with latent Markov models (Vermunt & Magidson, 2013), and with multilevel models (Bennink, Croon, & Vermunt, 2014), as well as with models combining these features, for example, a regression model for a continuous distal outcome in which not only the LC variable but also other predictors are included, or a structural equation model in which the LC variable is predicted by other variables and is itself a predictor of one or more distal outcomes. While all these possibilities exist and are available in software, there is a need for further research into the performance of the stepwise approaches in these more complex setups. For example, in models in which variance estimates are of interest, one should investigate the effect of the negative weighting used in BCH on the parameter estimates. It should be stressed that the LTB approach is more limited in the sense that it can only be used for the situation in which there is a single distal outcome, the situation investigated in this article. It should, however, be mentioned, that it can also be used with distal outcomes which are not continuous (Lanza et al., 2013). When used with categorical outcomes, the problems reported here do not occur and the LTB approach can be used without any concern.

Future research may focus on improving the recently proposed LTB method in various ways. First, it seems to be possible to prevent the encountered bias by expanding the logistic part of the model with quadratic and higher-order terms. Moreover, the undercoverage problem may be resolved by using better SE estimates, for instance, SEs obtained by bootstrapping. It may also be useful to transform the LTB approach into true stepwise approach, in which as in the three-step approaches the estimation of the LC model and the investigation of the association between classes and distal outcomes are fully separated. This would prevent the need to reestimate the original LC model for each distal outcome. Moreover, it would also make it possible to base the distal outcome analysis on a subsample, as was actually needed in our real data example, or even on a fully different sample, as would be the case when the classification information is obtained from an earlier study.

Another area that needs further attention are the somewhat underestimated BCH standard errors in the conditions with small sample size and low class separation. The SE estimates could be corrected for by accounting for the uncertainty in the BCH weights which are computed using the step-one parameter estimates, similarly to the correction proposed for the ML approach by Bakk et al. (2014). Another possible solution could be to switch to bootstrap SEs in these conditions. It should then be investigated whether bootstrapping in step three only suffices, or whether it is needed in step one as well.

A limitation of our study is that we focused on problems associated with heteroskedasticity and bimodality. It is recommended for future research to analyze whether other types of violations of normality, such as skewness, excess kurtosis, and outliers, are problem-

atic for the stepwise approaches at hand. We hypothesize that such violations will have only minor impact on the rather robust BCH approach, while they may bias parameter estimates of the ML and LTB methods to varying degrees.

Chapter 5

Relating latent class membership to continuous distal outcomes: improving the LTB approach and a modified three-step implementation

Abstract

The LTB approach relates latent classes (LCs) to distal outcomes by estimating a LC model with the outcome treated as covariate. Based on this model the class-specific means of the outcome are calculated. In this manner no distributional assumptions about the outcome are made Lanza et al. (2013). We provide a stepwise implementation of the approach that separates the building of the latent classes and the investigation of the relationship of the classes with the outcomes. Next, similar to quadratic discriminant analysis, we propose including a quadratic term in the logistic model for the LCs when the variances of the outcome are heteroskedastic in order to prevent parameter bias. And lastly we propose two alternative SE estimators (non-parametric bootstrap, jackknife), that yield better coverage rates than the currently used SE estimator proposed by Asparouhov and Muthén (2014) . The proposed improvements are tested via a simulation study with good results, and applied to real data.

This chapter is accepted for publication as Bakk, Z., Oberski, D.L. & Vermunt, J. K. Relating latent class membership to continuous distal outcomes: Improving the LTB approach and a modified three-step implementation. *Structural Equation Modeling*

5.1 Introduction

Latent class (LC) analysis is a well-known approach used in the social and behavioral sciences to create subgroups of units with similar scores on a set of observed indicator or response variables. In many applications, the interest lies not only in the clustering of units, but also in investigating whether LCs differ with respect to the mean of one or more continuous distal outcome variable. For example, De Cuyper et al. (2008) compared the class-specific means of job insecurity for psychological contract type clusters and Mulder et al. (2012) compared the means of juvenile offenders clusters on outcomes measuring recidivism. Other examples are predicting alcohol dependence from early substance abuse clusters and predicting the contraction of sexually transmitted diseases from sexual risk behavior clusters (Lanza et al., 2013).

The class-specific means of a continuous distal outcome can be estimated by expanding the LC model with the outcome as an additional indicator. The main problem with this approach, which is referred to as the one-step approach, is that assumptions have to be made about the within-class distribution of the distal outcome. Typically this will be the assumption of normality. However, in case this assumption is violated the whole LC solution can change when including the distal outcome, or even more classes can be extracted than would without this variable included (Bauer & Curran, 2003).

Lanza, Tan, and Bray propose an approach (called LTB approach after the developers) that bypasses the difficulties arising from potential violations of distributional assumptions. It involves estimating a LC model in which the distal outcome variable used as a covariate affecting the LCs instead of a response variable. Subsequently, using the estimates from this model, the class-specific means of the distal outcome variable are calculated (Lanza et al., 2013). The approach is implemented in the mainstream software for LC analysis, such as Mplus 7.1 (Muthén & Muthén, 1998-2012), and Latent GOLD 5.0 (Vermunt & Magidson, 2013).

While promising, the LTB approach has a few shortcomings that we address in this paper. First of all, when the distal outcome has heteroskedastic errors across classes, the LTB method may yield biased estimates of the class-specific means (Bakk & Vermunt, in press). We show how this bias can be prevented by including a quadratic term in the multinomial regression model for the classes. This is similar to what is done in a quadratic discriminant analysis. Furthermore while in the original article (Lanza et al., 2014) no standard error estimator was proposed, Asparouhov and Muthén (2014) proposed an ad-hoc estimator that is downward biased (Bakk & Vermunt, in press; Asparouhov & Muthén, 2014), thus obtaining too low coverage rates. We propose resolving this problem by using bootstrap-based or jackknifed standard errors.

Furthermore we propose a three-step estimation of the LTB approach. Many applied researchers prefer to first establish a measurement model, and in a later stage relate it to external variables of interest. It is also common that the measurement model is built by a researcher, and the structural model (relating LC membership to external variables) is built by different researchers. In this type of situations it is useful to have a three-step approach available. This proceeds as follows: 1) a standard LC analysis is performed using only the indicator variables, 2) individuals are assigned to latent classes, and 3) the assigned class scores are regressed on the distal outcome of interest, while correcting

for the classification error introduced in the second step (Vermunt, 2010). Based on the parameters obtained in the third step, the class-specific means of the distal outcome can be calculated. This three-step implementation can also be useful when the model estimates using the LTB approach is part of a larger, complex model.

In the remainder of this paper, we first introduce the basic LC model, then present the simultaneous LTB approach (as proposed by Lanza et al. 2014), and subsequently discuss its three-step implementation. Next, we introduce the proposed correction for the situation where the distal outcome has heteroskedastic errors, followed by the introduction of the alternative SE estimators. Subsequently, we present the results of a simulation study investigating the performance of the proposed improvements, and we demonstrate the use of the proposed methods via an example explaining respondent's income from parents' social status. Lastly, we conclude and suggest directions for future research.

5.2 The basic LC model

Let Y_{ik} denote the response of individual i on one of K categorical indicator variables, where $1 \leq k \leq K$ and $1 \leq i \leq N$. The full response vector is denoted by $\mathbf{Y}_i$. LC analysis assumes that respondents belong to one of the T categories of an underlying categorical latent variable X which affects the responses (McCutcheon, 1987; Goodman, 1974; Hagenaars, 1990). Denoting a particular latent class by t, the model can be formulated as follows:

$$P(\mathbf{Y}_i) = \sum_{t=1}^{T} P(X = t) P(\mathbf{Y}_i | X = t), \tag{5.1}$$

where $P(X = t)$ represents the (unconditional) probability of belonging to latent class t and $P(\mathbf{Y}_i|X = t)$ represents the class-specific response probabilities on the indicators. Furthermore, we assume that the K indicator variables are independent within classes, which is known as the local independence assumption. This yields:

$$P(\mathbf{Y}_i) = \sum_{t=1}^{T} P(X = t) \prod_{k=1}^{K} P(Y_{ik} | X = t). \tag{5.2}$$

For categorical responses, $P(Y_{ik}|X = t) = \prod_{r=1}^{R_k} \pi_{ktr}^{I(Y_{ik}=r)}$, where π_{ktr} is probability of response r on variable k for class t, and $I(Y_{ik} = r)$ is an indicator variable taking on the value 1 if $Y_{ik} = r$ and 0 otherwise.

The basic LC model can be extended to include a continuous distal outcome variable, which involves adding this variable to the model as an additional indicator and defining its class-specific distribution. However, this approach is hardly ever used in practice. Alternative approaches are the LTB approach and the three-step approach discussed in the next sections.

5.3 The simultaneous LTB approach

The LTB approach was developed with the goal to make it possible to estimate the association between the LC membership and the distal outcome without making strong distributional assumptions about the latter. This is especially important in case of a continuous outcome variable, in which case the assumption of normal class-specific distribution is often violated. As a consequence of violating this assumption a completely different LC model can be estimated, when the distal outcome is added, a possibility that is often not intended/ desired by researchers. Because of this issues the LTB approach is preferred over the one-step approach in many applications. Using the LTB approach, first a LC model is estimated with the distal outcome, say Z, included as covariate to the basic LC model. Subsequently, using Bayes theorem the class-specific means of the distal outcome are calculated (see Figure 1). We will call this approach originally proposed by Lanza, Tan, and Bray (2014) the simultaneous LTB approach.

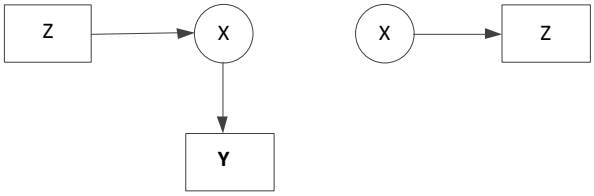

Figure 5.1: Original LTB approach

In step one, Z is included as a covariate to the basic model, by extending Equation 5.2 to model $P(\mathbf{Y}_i|Z_i)$ instead of $P(\mathbf{Y}_i)$ (Dayton & Macready, 1988; Bandeen-Roche et al., 1997):

$$P(\mathbf{Y}_i|Z_i) = \sum_{t=1}^{T} P(X = t|Z_i)) \prod_{k=1}^{K} P(Y_{ik}|X = t). \tag{5.3}$$

This model assumes that the indicator variables are conditionally independent of the covariate given the latent variable X. This is a standard assumption, made by the approaches available in literature to relate LC membership to external variables. If a direct effect is hypothesized, this should be explicitly modeled. The $P(X = t|Z_i)$ is parametrized using a multinomial logistic regression model:

$$P(X = t|Z_i) = \frac{e^{\alpha_t + \beta_t Z_i}}{1 + \sum_{t'=1}^{T-1} e^{\alpha_{t'} + \beta_{t'} Z_i}}, \tag{5.4}$$

where α_t and β_t are the intercept and slope coefficients for class t.

Next, in step two, the class-specific means μ_t are computed. These means equals:

$$\mu_t = \int_Z Z \; f(Z|X = t) \tag{5.5}$$

where $f(Z|X = t)$, the class-specific distribution of Z, is obtained as follows (Lanza et al., 2013):

$$f(Z|X = t) = \frac{f(Z)P(X = t|Z)}{P(X = t)}. \tag{5.6}$$

The quantities $P(X = t|Z)$ and $P(X = t)$ can be obtained from the estimated LC model. The distribution of Z, $f(Z)$, can be approximated using the empirical distribution of this variable (Asparouhov & Muthén, 2014). That is, replacing the integral in Equation 5.5 by a sum over the N sample units and replacing $f(Z)$ in Equation 5.6 by $\frac{1}{N}$ (Bakk & Vermunt, in press). This yields:

$$\mu_t = \sum_{i=1}^{N} Z_i \frac{P(X = t|Z_i)}{N \; P(X = t)}. \tag{5.7}$$

Simulation studies show that the estimated class-specific means obtained with this implementation of the LTB approach are unbiased as long as the relation between X and Z is linear-logistic (Asparouhov & Muthén, 2014; Bakk & Vermunt, in press). However, when the linearity does not hold the estimates are biased (Bakk & Vermunt, in press). This occurs for instance when the distal outcome has heteroskedastic errors. It turns out that larger the differences in the variances between classes, the larger the bias. We address this problem in more detail in section 4.

Lanza et al. (2013) did not discuss how to obtain SEs for the class-specific means, which are needed to make statistical inference possible. As a way out, Asparouhov and Muthén (2014) suggested obtaining approximate SEs by taking the square root of the within-class variance divided by the class-specific sample size; that is,

$$\sigma_t^2 = \sum_{i=1}^{N} (Z_i - \mu_t)^2 \frac{P(X = t|Z_i)}{NP(X = t)} \tag{5.8}$$

Simulation studies show that the Mplus approximate SE estimates underestimate the true sampling variability of the class-specific means (Asparouhov & Muthén, 2014; Bakk & Vermunt, in press), a problem that we address in section 5.

5.4 The three-step LTB approach

While originally proposed as a simultaneous estimation procedure, the LTB approach can easily be transformed into a three-step estimation procedure similar to the one proposed by Vermunt (2010). This can be beneficial mostly because it better follows the logic of researchers, who prefer to first establish a measurement model, and later associate it with one or more distal outcomes. Furthermore, it can be computationally less demanding

when the LTB approach is used with multiple distal outcome(s). The alternative is to repeat the full LTB analysis for every distal outcome, which means that larger, more complex models need to be estimated in all the different runs. Moreover, when there are missing values on the Z variables, also the sample may vary per distal outcome, which may yield additional differences in the definition of the latent classes. Finally, in some situations the simultaneous LTB approach cannot be used at all, for example, when the sample used to estimate the LC measurement model does not (fully) overlap with the sample containing the distal outcomes of interest (Bakk & Vermunt, in press).

The three-step LTB approach can be implemented as follows. Steps one and two involve performing a standard LC analysis (without distal outcome) and assigning individuals to classes, whereas in step three the assigned class memberships are related to the external variables of interest while correcting for classification errors, followed by the calculation of the class-specific means (Vermunt, 2010; Bakk et al., 2013) (see Figure 2). Using this approach, the first two steps need to be performed only once. Step three is repeated for each distal outcome variable, while keeping the measurement model parameters and the resulting classifications fixed.

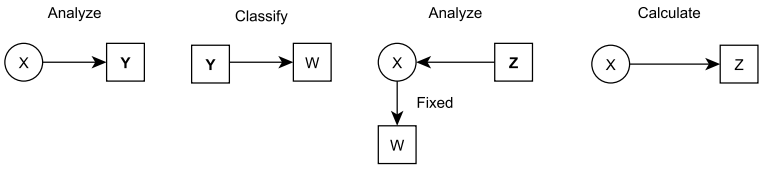

Figure 5.2: Three-step LTB approach

After estimating the step-one model (that includes only the indicator variables), as described in Equation 5.2, the units are assigned to the latent classes based on their posterior class membership probabilities: $P(X|\mathbf{Y}_i)$. During the assignment process a new variable, W_i is created, which equals the assigned class membership score for person i. Different assignment rules can be used, the best-known ones being modal and proportional assignment. Using modal assignment, each unit is assigned to a single class; namely, to the class for which the posterior membership probability is largest (Bolck et al., 2004), yielding what is called a hard partitioning. Using proportional assignment, each unit is assigned to each of the T classes with a weight equal to $P(W_i = s|\mathbf{Y}_i) = P(X = s|\mathbf{Y}_i)$, leading to what is sometimes referred to as a soft partitioning (Dias & Vermunt, 2008). Irrespective of the assignment rule used, there will be classification errors unless all classifications are perfect. These errors can be quantified as the off-diagonal elements of the $T - by - T$ classification table with entries $P(W_i = s|X = t)$ (Bolck et al., 2004; Vermunt, 2010; Bakk et al., 2013).

In step three, a LC model is estimated with W as a single indicator of class membership and with Z as a covariate affecting the classes:

$$P(W_i = s|Z_i) = \sum_{t=1}^{T} P(X = t|Z_i)P(W_i = s|X = t), \tag{5.9}$$

where $P(W_i = s|X = t)$ is fixed to the estimated values from step two, and $P(X = t|Z_i)$ contains the logistic parameters to be estimated. Next, just as with the simultaneous LTB approach, with the estimated values for $P(X = t|Z_i)$, the class-specific means of Z are calculated using Equation 5.7. This three-step LTB analysis is implemented in the Latent GOLD 5.0 program (Vermunt & Magidson, 2013).

5.5 The LTB approach with a quadratic term

While not stated explicitly by Lanza et al. (2013), if Z is normally distributed within classes with means μ_t and variances σ_t^2, the logistic model for $P(X = t|Z_i)$ is actually described by the discriminant function (Narsky & Porter, 2013, pp. 221-225). That is:

$$\log P(X = t|Z_i) = \log P(X = t) - \frac{1}{2}\log \sigma_t^2 - \frac{\mu_t^2}{2\sigma_t^2} + \frac{Z_i\mu_t}{\sigma_t^2} - \frac{Z_i^2}{2\sigma_t^2} + C,$$

where C is a constant ($-\frac{1}{2}\log(2\pi)$). This implies that $log(P(X = t|Z_i)$ is a quadratic function of Z:

$$\log P(X = t|Z_i) = a_t + b_t Z_i + c_t Z_i^2.$$

Where

$$\begin{aligned} a_t &= \log P(X = t) - \frac{1}{2}\log \sigma_t^2 - \frac{\mu_t^2}{2\sigma_t^2}, \\ b_t &= \frac{\mu_t}{\sigma_t^2}, \\ c_t &= -\frac{1}{2\sigma_t^2}. \end{aligned}$$

Thus, using the logistic formulation, the model for $P(X = t|Z_i)$ equals:

$$P(X = t|Z_i) = \frac{\exp(\alpha_t + \beta_t Z_i + \gamma_t Z_i^2)}{\sum_{t'=1}^{T}\exp(\alpha_{t'} + \beta_{t'} Z_i + \gamma_t Z_i^2)}. \tag{5.10}$$

In a multinomial logistic regression, one would impose identifying constraints on the α_t, β_t, and γ_t terms, for example, set them equal to 0 for class T. This means $\alpha_t = a_t - a_T$, $\beta_t = b_t - b_T$, and $\gamma_t = c_t - c_T$.

Since Z_i and Z_i^2 are correlated, the estimates of $P(X = t|Z_i)$ based on Equation 5.4 which does not contain the quadratic term and resulting estimates of the class-specific means will be biased unless the γ_t term is equal to 0. It should be noted that $\gamma_t = 0$ when the variances σ_t^2 are equal across classes, that is, when errors are heteroskedastic. However, when variances are unequal across classes, the quadratic term should be included in the multinomial logistic regression model to obtain the correct estimates for μ_t. By plugging in the estimates for $P(X = t|Z_i)$ obtained using Equation 5.10 into Equation 5.7, unbiased estimates of the class-specific means can also be obtained in the case of heteroskedastic errors.

5.6 Alternative SE estimators

Another issue with regard to the LTB approach implementation that needs further attention is the problem of the underestimated standard errors reported by Asparouhov and Muthén (2014) and Bakk and Vermunt (in press). This occurs because the Mplus approximate SEs do not take into account the sampling variability of the logistic parameters defining $P(X = t|Z_i)$ nor the fact that it conditions on the sample distribution of Z. A natural way to obtain SEs in such situations is by means of non-parametric resampling methods, that can be done by either a non-parametric bootstrapping or using the jackknife procedure.

5.6.1 Bootstrap SEs for the LTB approach

For the simultaneous LTB approach, bootstrap SEs (Guan, 2003) are obtained as follows:

1. Draw B random replication samples with replacement from the original data set.
2. Obtain the class-specific means of Z for each of these B bootstrap samples by applying the LTB approach.
3. Calculate the standard deviations of the class-specific means across the B bootstrap replications. This yields the bootstrap SE estimates.

For the three-step LTB method, a choice can be made whether to bootstrap only the third step or also the first step. In the latter case, one would also account for the uncertainty about the classification errors, which are fixed parameters in the step-three analysis. However, such a double bootstrap is much more costly and complex; that is, for each first-step bootstrap replication one should perform a full bootstrap of the third step. Because preliminary analyses showed that the step-three bootstrap is much more important for the SEs, we decided to bootstrap only the step-three parameters and evaluate the performance of this approach in the simulation study.

Bootstrapping the third step is similar to the non-parametric bootstrap described above. The main difference is that we sample from a data set with Z values and posterior class membership probabilities instead of Z values and Y values. That is:

1. Draw B random replication samples with replacement from the data set containing the distal outcome(s) of interest and the classification probabilities.
2. Obtain the class-specific means of Z for each of these B samples by applying the step-three LTB approach.
3. Calculate the standard deviation of the class-specific means across the B replications. This gives the bootstrap SE estimates.

5.6.2 Jackknife standard errors for the LTB approach

When using the jackknife approach, first the ML estimates of the parameters of interests (the class-specific means μ_t) are obtained based on the full sample of size N. Following, the estimates are recalculated leaving out one observation i at a time. The jackknife SE estimator is defined as follows:

$$SE(\mu_t) = \sqrt{\frac{N-1}{N}\sum_{i=1}^{N}(\hat{\mu}_t - \hat{\mu}_t(-i))^2}, \tag{5.11}$$

where $\hat{\mu}_t$ is the original estimate and $\hat{\mu}_t(-i)$ the estimate when leaving out observation i.

In the three-step approach, the jackknife estimator can be applied in the step-three analysis, when the parameter estimates pertaining to the $Z - X$ relationship and the corresponding $\hat{\mu}_t$ are obtained.

5.7 Simulation study

To evaluate the performance of the proposed adaptations of the LTB approach, we performed a simulation study. These adaptations are the inclusion of a quadratic term, the use of bootstrap and jackknife SEs, and the three-step variant of LTB approach.

In the simulation study we compare the performance of the LTB approach with the different modifications to the BCH approach. This is done because the BCH approach is known to be the most robust stepwise estimator for relating LC membership to continuous distal outcomes (Bakk & Vermunt, in press).

Data sets were generated from a four-class model for eight dichotomous indicators. The parameter settings were based on the LC application concerning psychological contract types described by Bakk et al. (2013). The class proportions were set to .50, .30, .10, and .10, similarly to those in this application. In class 1, the probability of a positive answer was set to .80 for all indicators, and in class 4 to .20. In class 2, it was set to .80 for the first four indicators and to .20 for the last four indicators, while in class 3 these settings were reversed.

The distal outcome variable Z was specified to have means of -1, -0.5, 0.5, and 1 in the four classes. The variance of Z was fixed to 1 in classes 1 and 4, but varied in classes 2 and 3. More specifically, in these two classes, the variance was set to 1, 4, 9, or 25, which corresponds to four different degrees of heteroskedasticity (none, small, medium, and large), and thus to different degrees of deviation from linearity of the logistic model for the association between Z and the classes.

The second factor that was varied was the sample size, which was specified to be either 500 or 1000. For all combinations of heteroskedasticity and sample size conditions, 500 simulation replications were used. The simulation were done using the computer programs R (Venables, Smith, & the R Core Team, 2013) and Latent GOLD 5.0 (Vermunt & Magidson, 2013).

The LTB approach was applied with and without the quadratic term, in both its simultaneous and its three-step form, where the three-step variant was used with either

Table 5.1: Bias in the estimates of class-specific means under eight simulation conditions, averaged over 500 replications and over the four classes.

	Condition: heteroskedasticity × sample size condition							
	None		Small		Medium		Large	
Estimator	500	1000	500	1000	500	1000	500	1000
LTB, linear model								
Modal	0.005	0.004	0.020	0.012	0.063	0.074	0.103	0.293
Proportional	0.003	0.005	0.022	0.014	0.094	0.110	0.150	-0.313
Simultaneous	-0.003	0.003	0.092	0.075	0.237	0.226	0.239	-0.222
LTB, quadratic model								
Modal	0.002	0.006	0.012	-0.002	0.013	0.003	-0.010	0.007
Proportional	0.006	0.007	0.010	-0.002	0.011	0.003	-0.013	-0.005
Simultaneous	0.001	0.000	0.003	-0.003	0.002	0.000	-0.003	0.000
BCH								
Modal	-0.005	-0.002	-0.008	-0.002	-0.006	0.002	-0.002	-0.005
Proportional	-0.006	-0.002	-0.009	-0.002	-0.007	0.003	-0.002	-0.005

modal or proportional class assignment. This yielded six different implementations of the LTB method. Each of these was combined with both the approximate SEs, jackknife SEs, and bootstrap-based SEs. For the bootstrap SEs we use $B = 1000$ bootstrap samples to obtain stable estimates. The BCH approach was applied with both modal and proportional assignment, using the sandwich standard error estimator, as proposed by Vermunt (2010).

The six different LTB implementations and two BCH implementations were compared with respect to parameter bias and relative efficiency. The efficiency of the LTB implementations was compared to proportional BCH, by dividing the simulation standard deviations of the estimates by those using BCH. Moreover, coverage rates of the 95% confidence intervals obtained with the different SE estimators were compared. In the following the results are presented averaged over the classes (a weighted average is used, with the class size as weight).

In Table 5.1 we show the bias in the estimates of the class-specific means under the different conditions, averaged over 500 replications and over the four classes. As the first three rows of Table 5.1 show the linear LTB is an unbiased estimator only when there is no heteroskedasticity. As heteroskedasticity increases the bias increases using this approach. This results hold for both the simultaneous and three step implementations. However using the quadratic term unbiased estimates of the class specific means are obtained also in the high heteroskedasticity conditions. Comparing the estimates obtained with the LTB approach using the quadratic model and BCH we can see that results are comparable. The bias is the lowest using the simultaneous approach, however the differences are negligible (only on the third decimals).

Next Table 5.2 shows the relative efficiency of the LTB estimators as compared to the BCH method. In almost all conditions the LTB estimators (when used with the correct model) are more efficient then BCH. Furthermore the simultaneous LTB is the most efficient estimator in all conditions. This results is expected, since in general simultaneous

Table 5.2: Relative efficiency compared with the proportional-assignment BCH estimator. Shown are simulation standard deviations of the estimates divided by those using BCH.

	Condition: heteroskedasticity × sample size							
	None		Small		Medium		Large	
Estimator	500	1000	500	1000	500	1000	500	1000
LTB, linear model vs. BCH								
Modal	1.024	0.976	1.251	1.308	1.819	2.262	2.258	3.523
Proportional	0.966	0.965	1.407	1.489	2.279	2.902	3.199	4.718
Simultaneous	0.942	0.848	1.958	2.298	2.671	3.251	3.099	3.825
LTB, quadratic model vs. BCH								
Modal	1.040	0.988	1.006	0.990	0.912	0.873	0.822	0.884
Proportional	0.983	0.976	0.958	0.981	0.898	0.866	0.780	0.852
Simultaneous	0.942	0.848	0.946	0.904	0.848	0.834	0.792	0.834

estimators are more efficient then stepwise estimators.

Following Table 5.3 shows the coverage rates obtained with the different estimators. When the correctly specified LTB approach is used with the approximate SEs the coverage rate is low (below 90%), even in the larger sample size conditions. The coverage rate obtained using the jackknife and bootstrap approaches is closer to the nominal 95%. The coverage rates obtained with the three-step LTB (with both modal and proportional assignment) is lower (between 90%- 96%) than the coverage using the simultaneous approach (between 94%- 96%). However when the sample size is large enough even with the three-step implementation the coverage rate with both the jackknife and bootstrap estimators is close to the nominal rate. In all conditions the bootstrap estimator is somewhat better than the jackknife, however the differences are very small. The coverage rates obtained using the BCH approach while close to the nominal 95% rate (between 90%- 95%), are somewhat smaller than the coverage obtained with the LTB approach using the bootstrap or jackknife SEs.

In summary, when the within-class errors of Z are heteroskedastic, the quadratic term should be used to obtain unbiased estimates of the class-specific means. The three-step LTB approaches perform as well as the original simultaneous approach with regard to bias however they are somewhat less efficient. The bootstrap and jackknife SEs yield coverage rates much closer to the nominal 95% rate than the Mplus approximate SEs. Furthermore the LTB apporach (both the simultaneous and teh three-step implementation) proved to be more efficient than the BCH approach.

5.8 An example application

We will now illustrate the different LTB implementations with an application using data from the 1976 and 1977 General Social Survey, a cross-sectional survey of the English-speaking, non institutionalized adult population of the U.S.A., conducted by the National Opinion Research Center (*GENERAL SOCIAL SURVEY 1976-1977*, 1977). We built a LC model for parents' social status using mother's education, father's education, and prestige

Table 5.3: Coverage of 95% confidence intervals under the eight conditions. Performance is shown for three difference standard error estimators for LTB.

	Condition: heteroskedasticity × sample size							
	None		Small		Medium		Large	
Estimator	500	1000	500	1000	500	1000	500	1000
LTB, linear model								
Modal								
Approximate	0.830	0.839	0.816	0.782	0.780	0.735	0.780	0.721
Jackknife	0.909	0.924	0.927	0.923	0.905	0.892	0.894	0.825
Bootstrap	0.909	0.927	0.932	0.927	0.920	0.913	0.923	0.846
Proportional								
Approximate	0.843	0.856	0.781	0.770	0.691	0.642	0.672	0.594
Jackknife	0.903	0.926	0.913	0.923	0.845	0.822	0.791	0.720
Bootstrap	0.903	0.926	0.919	0.933	0.874	0.847	0.812	0.748
Simultaneous								
Approximate	0.863	0.887	0.770	0.781	0.673	0.679	0.674	0.888
Jackknife	0.946	0.957	0.891	0.896	0.803	0.789	0.818	0.811
Bootstrap	0.961	0.961	0.914	0.910	0.824	0.806	0.829	0.824
LTB, quadratic model								
Modal								
Approximate	0.829	0.832	0.856	0.871	0.881	0.881	0.906	0.891
Jackknife	0.900	0.921	0.935	0.931	0.932	0.946	0.959	0.941
Bootstrap	0.906	0.921	0.936	0.936	0.940	0.947	0.962	0.942
Proportional								
Approximate	0.836	0.856	0.872	0.885	0.888	0.907	0.922	0.904
Jackknife	0.904	0.924	0.932	0.930	0.942	0.948	0.955	0.936
Bootstrap	0.903	0.924	0.931	0.931	0.943	0.955	0.963	0.941
Simultaneous								
Approximate	0.861	0.880	0.878	0.895	0.900	0.924	0.925	0.905
Jackknife	0.947	0.956	0.944	0.954	0.951	0.953	0.963	0.936
Bootstrap	0.954	0.965	0.952	0.956	0.959	0.959	0.967	0.944
BCH								
Modal	0.907	0.919	0.927	0.929	0.932	0.930	0.945	0.954
Proportional	0.898	0.908	0.917	0.921	0.929	0.930	0.941	0.952

of the father's job as indicators. Education was measured on a five-point scale ranging from 0 to 4, where 0 corresponds to 'lower than high school' and 4 to 'graduate'. Father's job prestige measured on a scale from 12 to 82 which recoded into three categories: low (12-36), medium (37-61), and high (62-82) prestige. As distal outcome variable we chose the real income of the respondent in thousand dollars increments.

In step 1, we fitted various LC models with the three indicators and selected the three-class model as best fitting model ($L^2 = 98.10$, $p = 0.65$, entropy $R^2 = 0.66$). The bivariate residuals were also small. The parameters of the three-class model are presented in Table 5.4. Class one, the largest class, comprises of respondents whose parents had a lower social status, while class 2 corresponds to medium, and class 3, the smallest class, to high social status of the parents. Note that in the step-one analysis the full sample of 3029 respondents was used by keeping also cases with missing values on one or more of the indicators in the analysis.

Next we related the respondent's income to the latent classes using the one-step approach in which income is an additional indicator and the LTB approach, both with and without accounting for possible heteroskedastic errors. The LTB approach was used with the original and three-step implementation. The estimated class-specific means obtained with the different approaches are presented in Table 5.5. The estimates obtained using the four different LTB approaches are very similar. They show that the income is highest among those respondents whose parents have the highest social status, and lowest for those whose parents have the lowest social status. However, the estimates obtained with the one-step approach (with equal or unequal variances) are very different, especially for class 3, which is the result of the fact that its definition changes drastically (for details, see Table C1 and C2 in the Appendix). Using unequal variances does also not solve the problem of completely changed class definitions, implying that the results of the one-step approach cannot be meaningfully interpreted in this application.

These results obtained with this application are in line with previous research. That is, in conditions where the sample size is large and the separation between the classes is good, the LTB approach obtains unbiased estimates, even without the quadratic term (Bakk & Vermunt, in press, Table 5). However, in the one-step approach, the class solution can drastically change to fit the distribution of the distal outcome, which is what happens in this example. Note that while the original LTB approach yields similar class-specific means of income as the three-step approach, the class proportions and the class-specific response probabilities on the indicators change somewhat (see Table C3 and C4 in the Appendix).

Table 5.6 presents the SE estimates obtained with the approximate jackknife, and bootstrap estimator for the four LTB approaches. The bootstrap and jackknife SE estimates are larger than the approximate estimates for both the original and three-step approaches with and without quadratic term. In this application, all SE estimators yield the same conclusion with regard to the significance of the income difference across classes.

Table 5.4: The 3-class model of parents social status. Class proportions and conditional response probabilities

	low	medium	high
Class Size	.69	.24	.07
Father's job status			
low	.47	.31	.05
medium	.53	.67	.46
high	.00	.02	.49
Mother's education			
lt high school	.83	.14	.15
high school	.16	.78	.44
junior college	.00	.03	.01
bachelor	.01	.04	.30
graduate	.00	.01	.10
Father's education			
lt high school	.95	.08	.01
high school	.05	.86	.12
junior college	.00	.00	.05
bachelor	.00	.05	.38
graduate	.00	.00	.43

Note.The sample consists of 3029 respondents

Table 5.5: Estimates of the class-specific means of income for the classes of parental social status obtained with the different methods

Method	Model	μ class 1	μ class 2	μ class 3
Simultaneous LTB	linear	25.37	36.49	44.16
Simultaneous LTB	quadratic	25.73	35.76	45.68
3-step LTB	linear	26.43	37.88	44.40
3-step LTB	quadratic	26.74	36.94	44.85
Standard 1-step	equal variances	25.36	36.62	62.59
Standard 1-step	unequal variances	21.33	26.98	69.73

Note. Sample size is 2767 obtained by excluding missing values on income

Table 5.6: SE estimates of the class-specific means of income for the different LTB approaches with the approximate and bootstrap SE estimators

	SE estimator	Original		3-step	
		linear	quadratic	linear	quadratic
Class1	approximate	0.90	0.58	0.55	0.44
	bootstrap	1.03	0.78	0.63	0.63
	jackknife	0.98	0.83	0.65	0.59
Class2	approximate	1.16	1.57	1.45	1.31
	bootstrap	1.20	1.52	1.51	1.76
	jackknife	1.26	2.39	1.46	1.43
Class3	approximate	2.62	3.42	2.81	3.04
	bootstrap	3.00	3.30	2.96	3.15
	jackknife	2.87	4.58	2.96	3.22

5.9 Discussion

The basic idea of the LTB approach is that LC membership can be related to distal outcome variables by inverting the relationship, that is, by regressing the LC membership on the outcomes of interest. Based on the parameter estimates of this "reversed" model, the class-specific means of the distal outcome are calculated. The main benefit of the LTB approach is that no strong distributional assumptions have to be made about the distal outcomes. Furthermore this approach provides a direct test of the overall association between the LC variable and the distal outcomes.

In this paper, we presented three possible improvements of the LTB approach applied to continuous distal outcomes. One of these is that we proposed incorporating the LTB approach into a three-step estimation procedure. This is especially useful if one wishes to relate the LC variable to multiple outcomes. It prevents the need to reestimate the full LC model with each distal outcome and makes sure that the class definitions remain the same across distal outcomes. Furthermore, contrary to the original approach, the three-step approach also can be applied when the step-one and step-three samples do not (fully) overlap. A disadvantage of the three-step approach is that when the uncertainty about the step one estimates is large (low entropy and/or sample size) the parameters are somewhat biased (Bakk et al., 2013; Vermunt, 2010; Asparouhov & Muthén, 2014). Furthermore this implementation is less efficient than the simultaneous LTB.

We showed that omission of the quadratic term in the logistic model for the LC variable yields biased estimates of the class-specific means when the distal outcome has heteroskedastic errors; that is, when means and variances are dependent. In such situations, unbiased estimates of the class-specific means can be obtained by including the quadratic effect of the distal outcome on the classes. We proposed using a jackknife or bootstrap-based SE estimator as an alternative to the currently used approximate estimator. Contrary to the latter, the bootstrap and jackknife estimators can account for the overall sampling variability in the distal outcome and for the sampling fluctuation in the

logistic parameters for the association between the LCs and the distal outcome.

The results of the simulation study showed that unbiased parameter estimates can be obtained if the heteroskedastic error is modeled using the quadratic term in the logistic regression of the LC variable on the outcome. This is the case with both the original and three-step implementation. Even in the highest heteroskedasticity conditions the LTB approach with the quadratic term is just as unbiased as the BCH approach, which is known to be the most robust estimator (Bakk & Vermunt, in press). At the same time the jackknife and bootstrap SEs obtain coverage rates close to the nominal 95% rate for both the original and three-step implementation of the LTB approach. The coverage rate is somewhat better using the simultaneous estimator than the three-step estimators. The LTB coverage rate (with both simultaneous and three-step implementation) is somewhat better than the one obtained using the BCH approach.

Furthermore we compared the relative efficiency of the different LTB estimators to the BCH approach. When used with the correct model the LTB estimators are more efficient than the BCH approach in all conditions. The simultaneous LTB is the most efficient estimator.

The real data example illustrated a situation where using a standard LC model with a distal outcome or the LTB approach makes a big difference. In the standard one-step approach, the full LC solution changed and became hard to interpret, while with the LTB approach this problem did not occur. Furthermore, the different implementations of LTB yielded similar results, which can be explained by the large sample size, the strong measurement model, and minor deviation from homoskedasticity.

We can conclude that the LTB method can be used with confidence for relating LC membership to continuous distal outcomes. However, attention should be paid to whether a linear or a quadratic model should be used. The results of the simulation study showed that when heteroskedasticity is present this effect should be added to the model. It is recommended to use the jackknife or the bootstrap standard errors in all conditions. The proposed three-step LTB can be used with a minimum loss of efficiency whenever using the original approach is less practical or not feasible and the uncertainty about the step one model is not too high.

It is recommended for further research to develop better tools for detecting whether a quadratic term needs to be added. For instance, the EPC-interest measure could be used, which quantifies how much the parameters of interest (here the class-specific means) change when adding the quadratic term (Oberski, 2014). Future research can also analyze the robustness of the LTB approach for other types of violations of possibly implicit assumption, such as when distal outcome distribution are skewed or show access kurtosis. In such situations, third- and fourth-order polynomials of the distal outcome might need to be included in the logistic model for the classes. However we do not assume that the impact of using higher than second order terms on the estimated class specific means will be important. It is also recommended to analyze its performance with continuous distal outcomes coming from other distributions than normal, such gamma, exponential, or beta distributions.

Chapter 6

Conclusions and discussion

In this dissertation we consider bias adjusted stepwise latent class analysis approaches that can be used to relate latent class (LC) membership to external variables of interest. Two main types of approaches are available: the bias-adjusted three-step approaches and the LTB approach. The three-step approaches first estimate the LC model with the indicator variables only, assign cases to classes, and relate the assigned scores in step three to external variables of interest while correcting for classification errors (Vermunt, 2010). Two implementations of the step-three model are available. The first one, the BCH approach estimates a model with observed variables only weighting the class assignment scores by the inverse of the classification error probabilities. Using the other approach -the ML approach- in step three a LC model is estimated with the class assignment as single indicator of LC membership with known misclassification probabilities. These three-step approaches were suggested for models with covariates only (Vermunt, 2010). The other option, the LTB approach, was developed specifically for situations where the LC membership is related to a distal outcome. Using this approach, first a LC model is estimated where the distal outcome is added as a covariate, and based on this model, the class-specific distribution of the distal outcome is calculated (Lanza, Tan and Bray, 2013). In the following, we discuss the different aspects with regard to the further development and testing of the three approaches described here that were addressed in this dissertation.

In Chapter 2, we proposed a generalization of the existing three-step methods. We show how the two existing three-step methods for latent class models with covariates can be generalized to a broader range of situations; that is, to formulate models for the joint probability of class membership and external variables. The correction methods can therefore now be applied in any situation where we wish to relate scores on class membership with external variables, irrespective of the hypothesized causal order. Though we focused mainly on the situation in which class membership is a predictor of a continuous, ordinal, or nominal outcome variable, the correction methods can be applied with distal outcome variables having any distribution from the exponential family. We also show how the ML correction method can be extended to models with more than one latent variable.

The performance of the correction methods was investigated by a simulation study and illustrated with two real data examples. The results of the simulation study show, similarly

to previously reported results, that the proposed three-step approaches yield unbiased estimates of the association between LC membership and external variables. All correction methods we tested performed well, meaning that their parameter estimates can be trusted. An exception occurs in the situations where the class separation of the measurement model is extremely low, in which case the step-three parameters are underestimated. Furthermore, the SE estimates show a small bias, which is more observable in situations with small sample sizes; with larger sample sizes, the SE bias is negligible.

In Chapter 3, we showed both analytically and by simulation that correct standard errors of the third-step model of the three-step approaches must incorporate the uncertainty about the classification error. We evaluatde eight possible types of standard error estimators for the ML approach, which we introduce based on the classic likelihood theory of Gong and Samaniego (1981). Although these standard error estimators are asymptotically equivalent under model correctness, they may yield different results in finite samples. A Monte Carlo simulation study showed that these SE corrections can make a large difference when the uncertainty about the first-step parameters is substantial. On the other hand, when the uncertainty about fixed estimates is low, the standard error corrections are not needed. Low uncertainty about the classification error will occur with large first-step sample sizes and high entropy R^2 (high class separation). No substantial differences between inference based on corrected versus uncorrected standard errors were found with first-step sample sizes above 2000 combined with entropy $R^2 > 0.90$. We also noted little difference between the asymptotic and finite-sample version of the corrected SEs. The asymptotic corrected standard error estimator (Oberski & Satorra, 2013), which is considerably easier to compute, is therefore recommended.

In Chapter 4, we investigated via a simulation study the robustness of the stepwise methods for models with continuous distal outcomes. We generated data under different degrees of heteroskedasticity and bimodality of the class-specific distributions of the distal outcome. Bimodality is a violation of the assumption of normality needed for the BCH and ML approaches; heteroskedasticity violates the assumption of logistic linearity of the LTB approach, and the assumption of homoskedasticity of the ML approach, when modeled with equal variances across classes. The simulation results revealed that the BCH method is the most robust approach: it yielded unbiased estimates under all investigated conditions. This is most probably the result of the fact that it involves performing a weighted ANOVA, a method that is known to be robust against violations of model assumptions. The other methods are sensitive to the violations considered. The ML approach fails to different degrees in all the situations investigated. When the variance is heteroskedastic, modeling it as equal between the classes produces a bias in the class-specific means. The ML approach cannot handle bimodal class-specific distributions of the outcome variable and probably any other departure from normality as well. The LTB approach yields biased estimates of the class-specific means when the errors are heteroskedastic, and shows a small bias with certain conditions of bimodality.

In Chapter 5, we presented three possible improvements of the LTB approach when applied to continuous distal outcomes. One of these is that we proposed incorporating the LTB approach into a three-step estimation framework. This is especially useful if one wishes to relate the LC variable to multiple outcomes. Furthermore, we showed that omission of the quadratic term in the logistic model for the LC variable yields biased estimates

of the class-specific means when the distal outcome has heteroskedastic errors; that is, when means and variances are not independent. In such situations, unbiased estimates of the class-specific means can be obtained by including the quadratic effect of the distal outcome on the classes. We also proposed using either a jackknife or a bootstrap-based SE estimator as an alternative to the currently used approximate estimator. Contrary to the latter, the bootstrap and jackknife estimators can account for the overall sampling variability in the distal outcome distribution and in the logistic parameters describing the association between the LCs and the distal outcome.

The results of the simulation study showed that unbiased parameter estimates can be obtained if the heteroskedastic error is modeled using the quadratic term in the logistic regression of the LC variable on the outcome. This applies to both the original and three-step implementation. At the same time, the jackknife and bootstrap SEs obtain coverage rates close to the nominal 95% rate for both the original and three-step implementation of the LTB approach. It should be mentioned that the coverage rate obtained with the jackknife estimator is slightly better than the one obtained by bootstrapping, though both are close to the nominal 95% level.

As a result of our study, the different developments discussed have been integrally implemented in the standard latent class analysis software Latent GOLD 5.00 (Vermunt & Magidson, 2013), making the methods developed here directly available to applied researchers. All three approaches are also available in Mplus (Muthén & Muthén, 1998-2012), however, the corrected SE estimators were not yet implemented in that software.

An important limitation of the three-step approaches presented/developed in this thesis is that in situations where the class separation is low the parameter estimates of the third step model are still downward biased. The reason for this is that in the first step the measurement model is too optimistic, yielding too low estimates for the classification errors. As such the correction term (which is based on the off-diagonal elements of the classification error matrix) is too low, thus leading to biased parameter estimates of the association of LC membership and external variables. In these situations, when the interest is on a distal outcome, the LTB approach might be the method of choice, because incorporating the external variables as covariates provides more information about the class membership while preventing the need of making strong distributional assumptions.

Three main types of further developments are recommended, which we shortly discuss in the following. First of all, further work is needed on the application of three-step approaches with more complex LC models. These approaches can be very useful to simplify larger complex models by estimating smaller pieces of the model separately, and combining the estimates in a step three model while controlling for classification error in the subparts. For example, the approach can be used for multilevel mixture models, where both at the lower and the higher level latent class models are defined. Such applications can be found for example in educational research, where clusters of students belonging to different ability clusters are nested in clusters of schools (for an example see Bennink, Croon, Keuning, and Vermunt (2014)). Estimating first a model at the lower level (students) and using the class assignments in a step-three model in which the level two model is estimated would simplify the analysis a lot. Other applications can be in mixture models for longitudinal data, for example, in latent Markov models. The measurement model could be established in step one, and the transitions over time can

be estimated separately in a step three model, where covariates affecting the transition probabilities can also be added. Work in this area was already started by Bartolucci, Montanari, and Pandolfi (2015) and Asparouhov and Muthén (2014).

There is need to further investigate the robustness of the three-step approaches under different types of violations of underlying model assumptions. In this thesis, we focused on violation of normality in case of continuous outcomes; however, the three-step approaches make more assumptions. For example, it should be investigated in more detail what the performance of the three-step approaches is when there are direct effects of covariates on the indicators, or when there are residual associations between a distal outcome and the indicators. Bennink, Croon, and Vermunt (2014) and Asparouhov and Muthén (2014) already started some work in this area, showing that when such direct effects are not taken into account, the step-three estimates can become biased.

Future work is also recommended with regard to the LTB approach. While in its current form it can deal with only a single distal outcome at a time, it may be expanded to multiple distal outcomes. Recently, Bray, Lanza, and Tan (2014) suggested applying the LTB approach with more complex LC models containing both covariates and a distal outcome. However, the authors did not provide SE estimates for such models, though our jackknife- and bootstrap-based estimates may also work in these situations. Furthermore, it needs further investigation the severity of the paramter bias in such complex models estimated in one run when some parts of the model are misspecified.

Appendix A

Latent GOLD syntax files for the examples in Ch. 2

A1. User defined Latent GOLD syntax for Example 1

```
options
    output parameters=first standarderrors estimatedvalues;
    variables
    dependent Political Modal, Religiosity Modal;
    independent socialClass;
    latent political nominal 3, Religiosity nominal 3;
    equations
    Religiosity <- 1 + SocialClass;
    Political <- 1 + Religiosity + SocialClass;
    PoliticalModal <- (D~wei) Political;
    Religiosity Modal <- (F~ wei) Religiosity;
    D = {0.854843 0.078100 0.067056
    0.036183 0.890474 0.073343
    0.022912 0.113239 0.863849};
    F = {0.970735 0.029264 0.000000
    0.037784 0.883258 0.078959
    0.000000 0.050674 0.949326};
```

A2. Automated Latent GOLD syntax for Example 2

```
.
    step3 ml modal;
    variables independent socialclass nominal coding=last;
    latent
    Political nominal posterior=(Cluster11 Cluster12 Cluster13),
    Religiosity nominal posterior=(Cluster21 Cluster22 Cluster23);
    equations
    Religiosity <- 1 + socialclass;
    Political <- 1 + Religiosity + socialclass;
```

Appendix B

Results of the simulation study separately for each condition for Ch. 3

Table B1: Comparison of the different variance estimators for low separation,for the 3 sample sizes separatly for one parameter, β_{13} for modal and proportional assignment

Final	Components		500			1000			2000		
	2^{nd} step	3^{rd} step	se	se/sd	coverage	se	se/sd	coverage	se	se/sd	coverage
					Modal						
$\mathbf{\Sigma}_3$		$\mathbf{\Sigma}_3^H$	0.17	0.83	0.90	0.12	0.97	0.94	0.08	1.03	0.96
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^H$	0.20	0.98	0.95	0.13	1.06	0.97	0.09	1.12	0.97
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^H$	0.20	0.99	0.95	0.13	1.08	0.97	0.09	1.14	0.97
$\mathbf{\Sigma}_3$		$\mathbf{\Sigma}_3^R$	0.18	0.87	0.91	0.12	0.98	0.94	0.08	1.04	0.97
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^R$	0.21	1.01	0.95	0.13	1.07	0.97	0.09	1.12	0.97
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^R$	0.21	1.01	0.95	0.13	1.09	0.97	0.09	1.14	0.97
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^R$	$\mathbf{\Sigma}_3^R$	0.21	1.05	0.95	0.13	1.08	0.97	0.09	1.12	0.97
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^R$	$\mathbf{\Sigma}_3^R$	0.22	1.06	0.96	0.14	1.10	0.97	0.09	1.15	0.97
					Proportional						
$\mathbf{\Sigma}_3$		$\mathbf{\Sigma}_3^H$	0.19	1.04	0.97	0.14	1.16	0.98	0.10	1.25	0.99
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^H$	0.21	1.14	0.98	0.15	1.24	0.98	0.10	1.33	0.99
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^H$	0.21	1.11	0.97	0.14	1.22	0.98	0.10	1.31	0.99
$\mathbf{\Sigma}_3$		$\mathbf{\Sigma}_3^R$	0.17	0.92	0.94	0.12	1.00	0.95	0.08	1.07	0.96
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^R$	0.19	1.04	0.95	0.13	1.10	0.97	0.09	1.17	0.98
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^R$	0.19	1.00	0.95	0.13	1.07	0.97	0.09	1.14	0.98
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^R$	$\mathbf{\Sigma}_3^R$	0.20	1.07	0.95	0.13	1.11	0.98	0.09	1.17	0.98
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^R$	$\mathbf{\Sigma}_3^R$	0.20	1.04	0.95	0.13	1.08	0.97	0.09	1.15	0.98

Note: $\mathbf{\Sigma}_3^*$ is the 1^{st} and $\mathbf{\Sigma}_3^{**}$ the 2^{nd} order correction, as defined in equation 17 and 18, and $\mathbf{\Sigma}^H$ and $\mathbf{\Sigma}^R$ are the Hessian based and robust estimators

Table B2: Comparison of the different variance estimators for high separation,for the 3 sample sizes separatly for one parameter, β_{13} for modal and proportional assignment

Final	Components		500			1000			2000		
	2^{nd} step	3^{rd} step	se	se/sd	coverage	se	se/sd	coverage	se	se/sd	coverage
					Modal						
$\mathbf{\Sigma}_3$		$\mathbf{\Sigma}_3^H$	0.14	0.93	0.95	0.10	1.08	0.97	0.07	1.09	0.96
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^H$	0.14	0.94	0.95	0.10	1.09	0.97	0.07	1.10	0.96
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^H$	0.15	0.95	0.95	0.10	1.10	0.98	0.07	1.11	0.96
$\mathbf{\Sigma}_3$		$\mathbf{\Sigma}_3^R$	0.14	0.94	0.94	0.10	1.10	0.97	0.07	1.09	0.96
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^R$	0.15	0.96	0.95	0.10	1.10	0.98	0.07	1.11	0.96
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^R$	0.15	0.96	0.95	0.10	1.10	0.96	0.07	1.11	0.96
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^R$	$\mathbf{\Sigma}_3^R$	0.15	0.96	0.95	0.10	1.10	0.97	0.07	1.11	0.96
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^R$	$\mathbf{\Sigma}_3^R$	0.15	0.96	0.95	0.10	1.10	0.98	0.07	1.11	0.96
					Proportional						
$\mathbf{\Sigma}_3$		$\mathbf{\Sigma}_3^H$	0.15	0.95	0.95	0.10	1.14	0.98	0.07	1.11	0.97
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^H$	0.15	0.96	0.95	0.10	1.15	0.98	0.07	1.15	0.97
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^H$	0.15	0.95	0.95	0.10	1.14	0.98	0.07	1.14	0.97
$\mathbf{\Sigma}_3$		$\mathbf{\Sigma}_3^R$	0.14	0.92	0.95	0.10	1.10	0.96	0.07	1.10	0.96
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^R$	0.14	0.93	0.95	0.10	1.10	0.96	0.07	1.11	0.96
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^H$	$\mathbf{\Sigma}_3^R$	0.14	0.92	0.95	0.10	1.10	0.96	0.07	1.10	0.96
$\mathbf{\Sigma}_3^*$	$\mathbf{\Sigma}_2^R$	$\mathbf{\Sigma}_3^R$	0.14	0.93	0.95	0.10	1.10	0.96	0.07	1.11	0.96
$\mathbf{\Sigma}_3^{**}$	$\mathbf{\Sigma}_2^R$	$\mathbf{\Sigma}_3^R$	0.14	0.92	0.95	0.10	1.10	0.96	0.07	1.10	0.96

Note: $\mathbf{\Sigma}_3^*$ is the 1^{st} and $\mathbf{\Sigma}_3^{**}$ the 2^{nd} order correction, as defined in equation 17 and 18, and $\mathbf{\Sigma}^H$ and $\mathbf{\Sigma}^R$ are the Hessian based and robust estimators

Appendix C

The modified 3 class model for parents social status with the different approaches presented in Ch. 4

Table C1: The modified 3-class model of parents social status obtained using the one-step approach without accounting for heteroskedasticity

Class Size	0.69	0.30	0.02
Father's job status			
low	0.47	0.23	0.20
medium	0.53	0.64	0.68
high	0.00	0.13	0.13
Mother's education			
lt high school	0.82	0.12	0.42
high school	0.17	0.72	0.42
junior college	0.00	0.03	0.00
bachelor	0.00	0.11	0.13
graduate	0.00	0.03	0.03
Father's education			
lt high school	0.93	0.08	0.52
high school	0.07	0.66	0.26
junior college	0.02	0.00	0.00
bachelor	0.00	0.13	0.10
graduate	0.00	0.11	0.13

Table C2: The modified 3-class model of parents social status obtained using the one-step approach with accounting for heteroskedasticity

Class Size	0.59	0.33	0.08
Father's job status			
low	0.48	0.33	0.05
medium	0.52	0.67	0.47
high	0.00	0.01	0.48
Mother's education			
lt high school	0.92	0.17	0.16
high school	0.08	0.75	0.47
junior college	0.00	0.03	0.01
bachelor	0.00	0.04	0.27
graduate	0.00	0.01	0.08
Father's education			
lt high school	0.96	0.32	0.01
high school	0.04	0.65	0.17
junior college	0.00	0.01	0.05
bachelor	0.00	0.03	0.37
graduate	0.00	0.00	0.40

Table C3: The modified 3-class model of parents social status obtained using the original LTB approach without accounting for heteroskedasticity

Class Size	0.59	0.33	0.08
Father's job status			
low	0.48	0.33	0.05
medium	0.52	0.67	0.47
high	0.00	0.01	0.48
Mother's education			
lt high school	0.92	0.17	0.16
high school	0.08	0.75	0.47
junior college	0.00	0.03	0.01
bachelor	0.00	0.04	0.27
graduate	0.00	0.01	0.08
Father's education			
lt high school	0.96	0.32	0.01
high school	0.04	0.65	0.17
junior college	0.00	0.01	0.05
bachelor	0.00	0.03	0.37
graduate	0.00	0.00	0.40

Table C4: The modified 3-class model of parents social status obtained using the original LTB approach with accounting for heteroskedasticity

Class Size	0.61	0.31	0.08
Father's job status			
low	0.47	0.33	0.06
medium	0.52	0.67	0.48
high	0.00	0.00	0.47
Mother's education			
lt high school	0.92	0.15	0.16
high school	0.08	0.78	0.47
junior college	0.00	0.03	0.01
bachelor	0.00	0.04	0.27
graduate	0.00	0.01	0.09
Father's education			
lt high school	0.95	0.31	0.02
high school	0.05	0.65	0.20
junior college	0.00	0.01	0.05
bachelor	0.00	0.03	0.36
graduate	0.00	0.00	0.38

Appendix D

Latent GOLD syntax for the example application of the LTB approach in Ch. 4

In this appendix, we present the LG 5.0 syntax used for the real data example. The following is the syntax for the original one-step LTB approach:

```
‘‘options
output
parameters=effect standarderrors=npbootstrap profile=LTB;
variables
dependent fatherjob nominal, mothereduc nominal, fathereduc nominal;
independent income;
latent
Cluster nominal 3;
equations
Cluster <- 1 + income;
fatherjob <- 1 + Cluster;
mothereduc <- 1 + Cluster;
fathereduc <- 1 + Cluster;’’
```

In the options, we indicate what output we would like to have. The command 'profile=LTB' yields class-specific means for the independent variables, as well as their standard errors. As can been, in this example, the chosen type of standard error estimator is the nonparametric bootstrap SE (npbootstrap). This can also be jackknife or standard.

The subsection 'variables' defines the indicators as dependent variables, the distal outcome as independent variable, and the latent class variable. The section 'equations' defines the model of interest. The heteroskedastic model is obtained by changing the first line of the equations into

```
‘‘ Cluster <- 1 + income + income * income;’’
```

When using a step-three approach, one first estimates a LC model and saves the classification information and other variables of interest to an output file. In the third step, this output file is used as the data set to be analyzed. The class assignments are related to the distal outcome as follows:

```
‘‘ options
step3 modal ml simultaneous;
output
parameters=effect standarderrors=npbootstrap profile=LTB;
variables
independent income;
latent Cluster nominal posterior = ( Cluster#1 Cluster#2 Cluster#3 );
equations
Cluster <- 1+ income;’’
```

The choice of the type of three-step approach is defined by the command line 'step3 modal ml simultaneous'. In the definition of the LC variable one specifies the variables in the data file containing the posterior classification probabilities from the first step: '(Cluster#1 Cluster#2 Cluster#3)'.

Bibliography

Agresti, A. (2002). *Categorical data analysis. Second edition*. John Willey and Sons, Inc., New Jersey.

Ahlquist, J. S., & Breunig, C. (2012). Model-based clustering and typologies in the social sciences. *Political Analysis*, *20*(1), 92–112.

Alwin, F., Duane. (2007). *Margins of error. A study of reliability in survey measurement*. New York, NY Willey.

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling*.

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014). Relating latent class assignments to external variables: standard errors for correct inference. *Political Analysis*, 520-540.

Bakk, Z., Tekle, F. T., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, *43*, 272-311.

Bakk, Z., & Vermunt, J. K. (in press). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling*.

Bandeen-Roche, K., Miglioretti, D. L., Zegger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92*, 1375-86.

Bartolucci, F., Montanari, E., & Pandolfi, S. (2015). Three-step estimation of latent markov models with covariates. *Computational Statistics and Data Analysis*, *83*, 287-301.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implication for overextraction of latent trajectory classes. *Psychological methods*, *8*(3), 338-363.

Beissinger, M. R. (2013). The semblance of democratic revolution: Coalitions in ukraine's orange revolution. *American Political Science Review*, *107*(03), 574–592. doi: 10.1017/S0003055413000294

Bennink, M., Croon, M., Keuning, M., & Vermunt, J. (2014). Measuring student ability, classifying schools, and detecting item bias at school level based on student level dichotomous items. *Journal of educational and behavioral statistics*, *39*(3), 180-201.

Bennink, M., Croon, M., & Vermunt, J. (2014). Stepwise latent class models for explaining group-level outcomes using discrete individual-level predictors [working paper]. Retrieved from `http://members.home.nl/jeroenvermunt/bennink2014a.pdf`

Blackwell, M., Honaker, J., & King, G. (2012). Multiple overimputation: a unified approach to measurement error in missing data [Working Paper]. Retrieved from `http://gking.harvard.edu/files/gking/files/measure.pdf`

Blaydes, L., & Linzer, D. A. (2008). The political economy of women's support for fundamentalist islam. *World Politics*, *60*, 579-609.

Bolck, A., Croon, M., & Hagenaars, J. A. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*, 3-27.

Bray, C. B., Lanza, S. T., & Tan, X. (2014). Eliminating bias in classify-analyze approaches for latent class analysis. *Structural Equation Modeling*, *online first*, 1-11.

Breen, R. (2000). Why is support for extreme parties underestimated by surveys? A latent class analysis. *British Journal of Political Science*, *30*, 375-82.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. (2006). *Measurement error in nonlinear models: A modern perspective, second edition*. New York: Chapman and Hall/CRC.

Chan, T. W., & Goldthorpe, J. H. (2007). Social stratification and cultural consumption: music in england. *European Sociological Review*, *23*, 1-19.

Clark, & Muthen, B. (2009). Relating latent class analysis results to variables not included in the analysis. *webnote*.

Clark, R. M., & Besterfield-Sacre, M. E. (2009). A new approach to hazardous materials transportation risk analysis: decision modeling to identify critical variables. *Risk Analysis*, *29*(3), 344-354.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). Wiley.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent class models. *Journal of the American Statistical Association*, *83*, 173-178.

De Cuyper, N., Rigotti, T., Witte, H. D., & Mohr, G. (2008). Balancing psychological contracts. Validation of a typology. *International Journal of Human Resource Management*, *19*, 543-561.

Dias, J. G., & Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics*, *23*, 643-59.

Feick, L. F. (1989). Latent class analysis of survey questions that include don't know responses. *Public Opinion Quarterly*, *53*, 525-47.

Feingold, A., Tiberio, S. S., & Capaldi, D. M. (2013). New approaches for examining associations with latent categorical variables: Applications to substance abuse and aggression. *Psychology of Addictive Behaviors*. Retrieved 2013-10-17, from `http://doi.apa.org/getdoi.cfm?doi=10.1037/a0031487` doi: 10.1037/a0031487

Fornell, C., & Larcker, D. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*, 39-50.

Fuller, A., Wayne. (1987). *Measurement error models*. New York, NY Willey.

General social survey 1976-1977. (1977). National Opinion Research Center. ICPSR ed. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, Conducted by University of Chicago. doi: doi.org/10.3886/ICPSR07398.v1

Glasgow, G., Golder, M., & Golder, S. N. (2012, April). New empirical strategies for the

study of parliamentary government formation. *Political Analysis*, *20*(2), 248–270. doi: 10.1093/pan/mpr058

Goetghebeur, E., Liinev, J., & Boelaert, M. (2000). Diagnostic test analyses in search of their gold standard: latent class analyses with random effects. *Statistical methods on medical research*, *9*, 231-248.

Gong, G., & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics*, 861–869.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology*, 79-259.

Grimmer, J. (2013). Appropriators not position takers: The distorting effects of electoral incentives on congressional representation. *American Journal of Political Science*, *57*, 624-642.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*, 267-297.

Guan, W. (2003). From the help desk: Bootstrapped standard errors. *The Stata Journal*(1), 71-80.

Gudicha, D. W., & Vermunt, J. K. (2011). Mixture model clustering with covariates using adjusted three-step approaches. In D. Lausen, van den Poel, & A. Ultsch (Eds.), *Algorithms from and for nature and life. Studies in classification, data analysis, and knowledge organization.*

Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators: local dependence models. *Sociological Methods and Research*, *16*, 379-405.

Hagenaars, J. A. (1990). *Categorical longitudinal data- loglinear analysis of panel, trend and cohort data*. Newbury Park, CA:Sage.

Hagenaars, J. A. (1993). *Loglinear models with latent variables*. Newbury Park, CA:Sage.

Hagenaars, J. A., & Halman, L. (1989). Searching for idealtypes. The potentialities of latent class analysis . *European Sociological Review*, *5*, 81-96.

Hill, J. L., & Kriesi, H. (2001). Classification by opinion-changing behavior: A mixture model approach. *Political Analysis*, *9*(4), 301–324.

Huang, & Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, *69*, 5-32.

Huang, Brecht, M.-L., Hara, M., & Hser, Y.-I. (2010). Influences of a covariate on growth mixture modeling. *J Drug Issues*, *40(1)*, 173-194.

Kauermann, G., & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, *96*(456), 1387-1396.

King, G., & Roberts, M. (2012). How robust standard errors expose methodological problems they do not fix [Working Paper]. Retrieved from `http://gking.harvard.edu/files/gking/files/robust.pdf`

König, T., Marbach, M., & Osnabrügge, M. (2013). Estimating party positions across countries and time. A dynamic latent variable model for manifesto data. *Political Analysis*, *21*(4), 468-491. doi: doi: 10.1093/pan/mpt003

Lanza, S. T., Tan, X., & Bray, C. B. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling*, *20:1*, 1-26.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Houghton Mill, Boston, MA.

Linzer, D. A. (2011). Reliable inference in highly stratified contingency tables: using latent class models as density estimators. *Political Analysis*, *19*, 173-187.

Loken, E. (2004). Using latent class analysis to model temperament types. *Multivariate Behavioral Research*, *39*, 625-652.

Marsh, H. W., Ludtke, O., Trautwein, U., & Morin, A. J. (2009). Classical latent profile analysis of academic self-concept dimensions: Synergy of person- and variable-centered approaches to theoretical models of self-concept. *Structural Equation Modeling*, *16*, 191-225.

McCutcheon, A. L. (1985). A latent class analysis of tolerance for nonconformity in the american public. *Public Opinion Quarterly*, *49*(4), 474–488.

McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, CA:Sage.

Mislevy, R. (1988). *Randomization-based inferences about latent variables from complex samples* (Tech. Rep.). Educational Testing Service.

Mulder, E., Vermunt, J. K., Brand, E., Bullens, R., & Van Marle, H. (2012). Recidivism in subgroups of serious juvenile offenders: Different profiles, different risks? *Criminal Behaviour and Mental Health*, *22*, 122-135.

Murphy, K. M., & Topel, R. H. (1985). Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics*, *3*, 88-97.

Mustillo, T. J. (2009). Modeling new party performance: A conceptual and methodological approach for volatile party systems. *Political Analysis*, *17*(3), 311-332.

Muthén, L. K., & Muthén, B. O. (1998-2012). Mplus users guide (7th ed.) [Computer software manual]. Los Angeles, CA: Muthén & Muthén.

Narsky, I., & Porter, F. C. (2013). *Statistical analysis techniques in particle physics: Fits, density estimation and supervised learning*. John Willey and Sons, Inc., New JerseyC.

Nylund, K. L., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*, 535-569.

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*(1), 45-60.

Oberski, D. L., & Satorra, A. (2013). Measurement error models with uncertainty about the error variance. *Structural Equation Modeling*, *20*.

Oberski, D. L., & Vermunt, J. K. (2013). A model-based approach to goodness-of-fit evaluation in item response theory. *Measurement: Interdisciplinary Research & Perspectives*, *11*, 117-122. doi: 10.1080/15366367.2013.835195

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, *46*(1), 27–29.

Olino, T. M., Lopez-Duran, N. L., Kovacs, M., George, C. J., Gentzler, A. L., & Shaw, D. S. (2011). Developmental trajectories of positive and negative affect in children at high and low familial risk for depressive disorder . *The journal of child psychology and psychiatry*, *52(7)*, 792-799.

Parke, W. R. (1986). Pseudo maximum likelihood estimation: the assymptotic distribution. *The Annals of Statistics*, *14*, 355-357.

Pastor, D. A., Barron, K. E., Miller, B. J., & Davis, S. L. (2007). A latent profile analysis of college students achievement goal orientation. *Contemporary Educational Psychology*, *32*, 847.

Petersen, J., Bandeen-Roche, K., Budtz-Jrgensen, E., & Groes, L. K. (2012). Predicting latent class scores for subsequent analysis. *Psychometrica*, *77(2)*, 244-262.

Petras, H., & Masyn, K. (2010). General growth mixture analysis with antecedents and consequences of change. In A. Piquero & D. Weisburd (Eds.), (p. 69-100). Springer, New York.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2001). Maximum likelihood estimation of generalized linear models with covariate measurement error. *The Stata Journal*.

Ristei Gugiu, M., & Centellas, M. (2013, May). The democracy cluster classification index. *Political Analysis*, *21*(3), 334–349. doi: 10.1093/pan/mpt004

Rubin, D. B. (1987). *Multiple imputation for nonresponse in survey research*. New York, NY Willey.

Schafer, J. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333-343.

Skinner, C., Holt, D., & Smith, T. (1989). *Analysis of complex surveys*. New York, Wiley.

Skrondal, A., & Kuha, J. (2012). Improved regression calibration. *Psychometrika*, *77*, 649-669.

Sniderman, P. M., Tetlock, P. E., Glaser, J. M., Green, D. P., & Hout, M. (1989, January). Principled tolerance and the american mass public. *British Journal of Political Science*, *19*(01), 25. doi: 10.1017/S0007123400005305

Stouffer, S. A. (1955). *Communism, conformity, and civil liberties: A cross-section of the nation speaks its mind*. Transaction Books.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528–540.

Tofighi, D., & Enders, C. K. (2008). *Identifying the correct number of classes in growth mixture models*. (G. R. Hancock & K. M. Samuelsen, Eds.). Charlotte, NC: Information Age Publishing.

Treier, S., & Jackman, S. (2008). Democracy as a latent variable. *American Journal of Political Science*, *52*(1), 201–217.

Van der Heijden, P., 't Hart, H., & Dessens, J. (1997). A parametric bootstrap procedure to perform statistical tests in a LCA of anti-social behaviour. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences*. (p. 196-208). New York, NY: Waxmann.

van der Heijden, P. G. M., Dessens, J., & Bockenholt, U. (1996). Estimating the concomitant-variable latent-class model with the em algorithm. *Journal of Educational and Behavioral Statistics*, *21*, 215-229.

Venables, W. N., Smith, D. M., & the R Core Team. (2013, April). *An introduction to R. Notes on R: A programming environment for data analysis and graphics. Version 3.0.0.* Retrieved from `http://cran.r-project.org/doc/manuals/R-intro.pdf`

Vermunt, J. K. (2010). Latent class mmodeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*, 450-469.

Vermunt, J. K., & Magidson, J. (2013). Technical guide for Latent Gold 5.0: Basic and advanced [Computer software manual]. Belmont Massachusetts:Statistical Innovations Inc.

Wang, C.-P., Brown, H. C., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association*, *100*, 1054-10.

Wedel, M., Ter Hofstede, F., & Steenkamp, J.-B. E. (1998). Mixture model analysis of complex samples. *Journal of Classification*, *15*, 225-244.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, *50*(1), 1-25.

Yamaguchi, K. (2000). Multinomial logit latent-class regression models: An analysis of the predictors of gender-role attitudes among japanese women. *American Journal of Sociology*, *105*, 1702-40.

Summary

This thesis focuses on the development of stepwise latent class analysis approaches for situations where LC membership is associated with external variables of interest. Two types of approaches are considered: the three-step approaches and the LTB approach.

Using the three-step approaches first a LC model is estimated with the indicator variables only, then in step two individuals are assigned to latent classes, and subsequently in step three the class assignments are used in further analysis while accounting for the measurement error introduced in step two. While until recently the three-step approach was used without correcting for the classification error introduced in step two, in this thesis we advocate for the use of bias-corrected step three models, based on the work of Bolck, Croon, and Hagenaars (2004), who showed how the amount error introduced in step two can be accounted for in step three. This work was further developed by Vermunt (2010), who introduced a user friendly implementation of the correction, and suggested a more direct approach, that we call the ML approach. Both approaches use the same classification error matrix for applying the bias adjustment. The BCH approach uses a model with observed variables only, weighting the LC membership by the inverse of the classification errors. The ML approach involves estimating a LC model in step three in which the class assignments are used as an indicator with known classification errors.

An alternative is the LTB approach (so called after the developers, Lanza, Tan and Bray, (2013)) that is specifically developed for models where LC membership is related to a distal outcome. Using this approach first an LC model is estimated with the distal outcome added as a covariate to the model, and in the next step the class-specific means of the distal outcome are calculated.

This thesis proposes various extensions of the bias-adjusted three-step approaches and the LTB approach. In Chapter 2, we show how the three-step approaches, which were initially developed only for models with covariates, can be extended to models with distal outcomes and multiple latent variables, making the applicability of the approaches more widespread. We show that when the model assumptions hold, both the ML and BCH approaches yield unbiased estimators of the relationship between LC membership and external variables in all special cases investigated.

In Chapter 3, we show that the currently available standard error estimators for the ML approach are biased and we suggest alternative SE estimators. More specifically, we show both analytically and by simulation that in the estimation of the standard errors of the third step estimator, one should incorporate the uncertainty about the classification error. We provide alternative standard error formulae based on the classic likelihood theory of

Gong and Samaniego (1981).

In Chapter 4, we investigated the robustness of the different stepwise LC analysis approaches to violations of underlying model assumptions when used with continuous distal outcomes. The BCH method, the ML method with equal variances, and the ML method with unequal variances assume that the distal outcome is normally distributed, whereas ML(equal) also assumes homoskedasticity. The LTB method assumes that the relationship between the distal outcome and class membership is linear on a logistic scale. The BCH approach turned out to be the most robust. The ML approach, which involves estimating a LC model is sensitive to model misspecifications, such as assuming equal variances across classes when this does not hold, and to nonnormality. At the same time, the LTB approach proved to be sensitive to violations as well; that is, estimates are biased when the error term is heteroskedastic.

Finally, in Chapter 5, we propose various improvements to the LTB approach. We show that using a quadratic term in the first-step logistic model relating the LC membership to the distal outcome, eliminates the bias in the class-specific means calculated from this model in situations where the error term is heteroskedastic. Furthermore, we propose using either a bootstrap or jackknife SE estimator, both of which perform better than the currently available estimator. We also suggest a true stepwise implementation of the procedure; that is, analyzing a step one model with indicators only, saving the class assignments, and running a step three model with the ML approach using the distal outcome as covariate. The class-specific mean of the distal outcome can be computed based on the step-three model parameters.

Acknowledgments

Hereby I would like to say a word of thanks to everyone who helped me in finalizing this dissertation. It was an amazing journey. During the years spent at Tilburg University I developed professionally and traveled the world- both much more then I could have ever imagined.

First of all I would like to thank my promotor, Jeroen Vermunt, and co-promoter Daniel Oberski. Jeroen, you were always available when I needed help, while giving me enough freedom to figure out my own ways. You had enough patience to teach me statistics from the basics in my master thesis project, till completing a thesis with 4 published articles. Daniel, you always brought in challenging new perspectives, that extended my horizons. Both of you, thank you.

Thanks goes out to the members of my defense committee as well, who gave valuable feedback on the thesis. Furthermore I would like to thank Fetene Tekle for his contributions to the second chapter. I am also thankful to the members of the Vici group- your valuable feedbacks shaped this thesis. Lianne, your comments and corrections on the dissertation were an enormous help. Special thanks goes out to Prof Le Roux at Stellenbosch University, and the whole department of Statistics, you made my exchange at Stellenbosch one of the nicest memories of my PhD years. The IOPS courses and conferences also shaped this period. Jacques, thanks for all that I learned from you in St Petersburg, and for all the beers.

My time at Tilburg was made special by my colleges, the shared lunches, coffees and chats. Special thanks to my roomiees: Margot, Daniel, Marie-Anne, Davide- it was so much fun chatting about science, and making geeky jokes with you. I am thankful for the friendships formed with you in this period. I am also thankful to Marieke and Liesbeth for all the administrative support, and good words. While I do not have enough space to mention all colleges by name, I am grateful to all of the MTO gang for the truly gezellig atmosphere. Dank jullie wel!

I am also most thankful to my family and friends who were always there for me during this period. Anyuci köszönöm, hogy végig támogattál, még ha nehéz is volt ilyen távol látni gyermeked. Zsofikam, Zsuzsi köszönöm a mindennapos türelmeteket, és támogatásotokat- nélkületek sokszor feladtam volna. Adriana si Zsofi mersi mult câ sunteti paranimfi atât de buni. Bianca, Magdi, Andrey, Ruth, thanks for all the great time spent with you in person and virtually, in- and outside Europe.